

EQUALITY AND LOAN ORIGINATION IN MORTGAGE LENDING: A COLLECTION OF
ESSAYS

by

LU FANG

(Under the Direction of Henry J. Munneke)

ABSTRACT

This dissertation focuses on the issue of whether inequality, specifically, gender inequality and racial inequality, exists in mortgage loan pricing. Prior studies on this inequality issue normally ignore the possibility that a borrower's personal traits including gender and race might be associated with loan termination patterns, specifically default patterns and prepayment patterns. Failure to account for the possible association between a borrower's personal traits and loan termination probability might lead to biased estimation of the impact of a borrower's gender or race on a loan's contract rate. Unlike prior studies, this dissertation examines whether a borrower's gender or race impacts the loan contract rate beyond the extent to which it affects the probability of a borrower defaulting upon or prepaying a loan. Using data on the subsequent performance of each loan in the sample, a competing-risks loan hazard model is employed to investigate the degree to which a borrower's gender or race is associated with loan default probability and prepayment probability. The resulting loan-level predicted default probability and prepayment probability are incorporated in a loan contract rate determination model to test whether a borrower's gender or race has additional impacts on contract rate. The first essay focuses on gender inequality, and the second one concentrates on racial inequality.

The results reveal that a borrower's personal traits are indeed associated with loan termination probability. In terms of a borrower's gender, the results show a female sole borrower tends to be more likely to default than borrowers in other gender groups; while with respect to race, African American borrowers and Hispanic borrowers are shown to be less likely to prepay than non-Hispanic White counterparts. After these associations are completely controlled for, the dissertation has found empirical evidence of both gender inequality and racial inequality in mortgage lending. Female sole borrowers are shown to pay higher contract rates than borrowers in other gender groups, while both African American borrowers and Hispanic borrowers tend to pay more for their mortgage loans than non-Hispanic Whites.

INDEX WORDS: Gender Equality, Racial Equality, Default Risk, Prepayment Risk, Loan
Pricing

EQUALITY AND LOAN ORIGINATION IN MORTGAGE LENDING: A COLLECTION OF
ESSAYS

by

LU FANG

B.S., Nanjing University, P.R. China, 2009

B.A., Nanjing University, P.R. China, 2009

M.A., Renmin University of China, P.R. China, 2011

A Dissertation Submitted to the Graduate Faculty of The University of Georgia in Partial
Fulfillment of the Requirements for the Degree

DOCTOR OF PHILOSOPHY

ATHENS, GEORGIA

2016

© 2016

Lu Fang

All Rights Reserved

EQUALITY AND LOAN ORIGINATION IN MORTGAGE LENDING: A COLLECTION OF
ESSAYS

by

LU FANG

Major Professor:	Henry J. Munneke
Committee:	James B. Kau
	Richard W. Martin
	Darren K. Hayunga
	James N. Conklin

Electronic Version Approved:

Suzanne Barbour
Dean of the Graduate School
The University of Georgia
August 2016

DEDICATION

This dissertation is dedicated to my beloved parents, for their unconditional love, support, and encouragement.

ACKNOWLEDGEMENTS

I especially would like to express my sincere and deep gratitude to Henry Munneke, my advisor and “academic farther”, for his enormous support, effort, and encouragement. I am also especially grateful to James Kau, my committee member and co-author, for his wisdom, encouragement, and guidance. I also much appreciate the help, constructive comments, and suggestions by other members of my dissertation committee, Richard Martin, Darren Hayunga, and James Conklin. Additionally, I also feel extremely fortunate to have known Carolyn Dehring and Michael Ericksen, and owe a debt of gratitude to them for their valuable mentoring and help during the first few years of my Ph.D. life. Besides, I would like to extend my gratitude to Kiplan Womack, Joseph Nicholson, Elizabeth Fisher, and Stephen Buschbom, for their kind and great assistance throughout the five years at Athens. I would also like to extend my heartfelt gratitude to all the faculties and staffs at the ILSRE Department, for their nice help and support. Last, but not least, my deepest gratitude and love goes to my parents, for their unconditional love, support, and encouragement.

TABLE OF CONTENTS

	Page
ACKNOWLEDGEMENTS	v
LIST OF TABLES	viii
CHAPTER	
1 INTRODUCTION	1
2 ESSAY ONE: GENDER EQUALITY IN MORTGAGE LENDING	5
Abstract	6
2.1. Introduction.....	7
2.2. Literature Review.....	10
2.3. Model	13
2.4. Data	19
2.5. Model Specifications	24
2.6. Results.....	29
2.7. Discussion	37
2.8. Conclusion	39
References.....	41
Appendices.....	51
3 ESSAY TWO: DON'T FORGET ABOUT DEFAULT AND PREPAYMENT – AN ANALYSIS OF MORTGAGE LENDING	54
Abstract	55

3.1. Introduction.....	56
3.2. Literature Review.....	61
3.3. Model	65
3.4. Data	75
3.5. Model Specifications	82
3.6. Results.....	90
3.7. Discussion	100
3.8. Conclusion	103
References.....	106
Appendices.....	117
4 CONCLUSIONS.....	124

LIST OF TABLES

	Page
Table 2.1: Descriptive Statistics of Mortgage Loans at Loan Origination	44
Table 2.2: Descriptive Statistics of Mortgage Loans at Loan Origination by Gender	45
Table 2.3: Loan Hazard Model Estimates.....	46
Table 2.4: Loan Contract Rate Estimates -2SLS	47
Table 2.5: Loan Contract Rate Estimates with Gender-Race Interaction -2SLS.....	48
Table 2.6: Nonparametric Matching Results on Nearest 1-to-1 Matching with Replacements and with Calipers	49
Table 3.1: Descriptive Statistics of Mortgage Loans at Loan Origination	109
Table 3.2: Descriptive Statistics of Mortgage Loans at Loan Origination by Race	110
Table 3.3: Default Hazard Model Estimates.....	111
Table 3.4: Prepay Hazard Model Estimates.....	112
Table 3.5: Loan Contract Rate Estimates on the Pooled Sample -2SLS	113
Table 3.6: Matching Results on Nearest 1-to-1 Matching between African Americans and Non- Hispanic Whites	114
Table 3.7: Matching Results on Nearest 1-to-1 Matching between Hispanics and Non-Hispanic Whites	115
Table 3.8: Loan Contract Rate Estimates on the Pooled Sample in a Race-Blind World – 2SLS	116
Table A.1: Definitions of Variables.....	51

Table A.2: Loan Contract Rate Reduced-Form Estimates.....	53
Table B.1: Definition of Variables.....	117
Table B.2: Matching Balance Diagnosis on Matching Sample 1 between African Americans and Non-Hispanic Whites.....	119
Table B.3: Matching Balance Diagnosis on Matching Sample 2 between African Americans and Non-Hispanic Whites.....	120
Table B.4: Matching Balance Diagnosis on Matching Sample 1 between Hispanics and Non- Hispanic Whites	121
Table B.5: Matching Balance Diagnosis on Matching Sample 2 between Hispanics and Non- Hispanic Whites	122
Table B.6: Matching Balance Diagnosis on Matching Sample 3 between Hispanics and Non- Hispanic Whites	123

CHAPTER 1

INTRODUCTION

Since the Equal Credit Opportunity Act (ECOA) was enacted in 1974, the issue of whether equality has been achieved in mortgage lending has drawn much attention, both politically and socially, and with a voluminous literature. Inequality might exist in mortgage markets in several forms. It may exist at the loan application stage. Mortgage lenders may treat loan applicants differently on the basis of an applicant's personal traits (including gender, race, and national original, etc.) by denying loan applications more frequently by a specific group of applicants. Inequality may also occur during the loan origination stage in which borrowers in a specific group are charged at a higher price for their approved mortgage loan applications. While earlier studies on the issue of inequality in mortgage lending concentrate more on this issue at the loan application stage, recent studies have shifted their focus to the loan origination stage. This shift is consistent with the evolution of the mortgage lending industry. Since the late 1990s with the development of the subprime mortgage markets, lenders started to provide credit to risky borrowers who they would have denied, but at a higher interest rate. This dissertation research, in accordance with this shift, also aims to evaluate the existence of inequality at the loan origination stage. In other words, the goal of this dissertation research is to investigate if a borrower's personal traits, specifically, a borrower's gender and race, affect loan contract rate.

Though numerous prior studies have analyzed this study question, one common potential problem involved in those prior studies is they ignore the possibility that a borrower's personal traits might be associated with loan termination risk including default risk and prepayment risk.

Previous studies commonly applied a reduced-form model to investigate the relationship between a borrower's personal traits and the loan contract rate. In a reduced-form model, a loan's contract rate is regressed against a borrower's personal traits as well as a series of covariates that are believed to impact loan termination risk, including the characteristics of the loan, the borrower, the collateral, and the neighborhood. The potential problem of this reduced-form approach is if some loan termination risk factors are unobserved in the data and are correlated with a borrower's personal traits, the estimation by a reduced-form model would be biased. In order to solve this potential problem, loan default probability and prepayment probability are explicitly modelled and accounted for in this dissertation research. A competing-risks loan hazard model is employed to investigate the relationship between a borrower's personal traits (gender and race) and loan default probability and prepayment probability. The loan-level predicted default probability and prepayment probability generated by this loan hazard model are accounted for in a loan contract rate determination model to examine whether a borrower's gender/race has an additional impact on loan contract rate, the impact that could not be explained by any financial reasons. Additionally, in terms of loan termination risk, different from prior studies that only focus on the impact of loan default risk on loan pricing, this dissertation research models both default risk and prepayment risk simultaneously through the competing-risks loan hazard model. The main reason for considering both default and prepayment risk is they interact with each other in a competitive way, therefore, a loan's value could not be priced adequately without consideration of any one of them (Kau et al., 1995).

This dissertation research also contributes to the existing literature by proposing and employing a non-parametric approach to analyze the relationship between a borrower's gender/race and contract rate that has not been used in previous studies. Specifically, matching is

conducted to pair a loan by a borrower in a particular gender/race group to a comparable loan by a borrower in another group with the closest distance on the predicted default probability and prepayment probability generated by the loan hazard model. This matching approach is aimed to eliminate any systematic differences in loan termination risk between the resulting two gender/race groups, and thus to test if there is any significant difference in loan contract rate between these two groups. This non-parametric matching approach is adopted as it has several advantages over the parametric regression technique. First, it does not depend on the specifications of the contract rate determination model. Additionally, it could generate a matched sample consisting of relatively more homogenous loans in terms of loan termination risk across gender/race groups for the purpose of comparison. However, one disadvantage of this non-parametric technique is it reduces the sample size. In this dissertation research, both of these two approaches are applied, and it is anticipated that they yield similar results.

Unlike prior studies that only concern whether inequality exists in mortgage lending, this dissertation research is not only interested in testing if a borrower's personal traits impact loan contract rate, but also interested in figuring out the potential sources of the inequality, if empirical evidence of inequality is found. Several potential explanations for the phenomenon of inequality are proposed, and tested when the data needed is available.

The goals of this dissertation research are addressed in two essays. The first essay focuses on the issue of gender inequality by examining if a borrower's gender affects loan contract rate beyond the extent to which it impacts loan default probability and prepayment probability. The second essay aims to evaluate the existence of racial (ethnic) inequality in mortgage loan pricing. In the second essay, two forms of racial (ethnic) inequality are investigated, the one based on a

borrower's race (ethnicity), and the other one on the basis of the racial (ethnic) composition of a borrower's neighborhood, commonly referred as redlining.

The dissertation is organized as follows. Chapter one provides a brief introduction of this dissertation research. Chapter two presents the first essay on gender inequality in mortgage lending. Chapter three demonstrates the second essay on racial inequality in loan pricing. The last chapter summarizes all the results of the two essays.

CHAPTER 2

ESSAY ONE: GENDER EQUALITY IN MORTGAGE LENDING¹

¹ Fang, Lu and Henry J. Munneke. To be submitted to *The Review of Financial Studies*.

Abstract

Few papers have attempted to empirically examine whether gender inequality exists in the mortgage market, even though this issue has been extensively explored in other markets. Using a sample of 30-year fixed-rate subprime mortgage loans, this paper empirically tests whether a borrower's gender affects the loan contract rate charged, more specifically if a borrower's gender impacts the loan contract rate beyond the extent to which it impacts the borrower's probability of default or prepayment. A loan hazard model is estimated to investigate the degree to which a borrower's gender is associated with loan default probability and prepayment probability using data on the outcome of each loan since origination. A loan contract rate model is then estimated to examine whether gender impacts the loan contract rate when the probability of default and prepayment are accounted for. The results reveal that borrowers of different gender have different loan termination patterns. In addition, this study indicates that after controlling for the correlation between a borrower's gender and the probability of a borrower defaulting or prepaying, female borrowers are shown to pay a higher contract rate in the subprime mortgage market over the study period.

2.1. Introduction

Research related to gender equality in employment opportunities, education, housing, and business activities has become more prevalent in the literature over the last 30 years. For instance, the issue of a gender pay gap has been extensively explored for decades, and continues to draw a lot of attention (Altonji & Blank, 1999; Marianne, 2011).² There have also been intense debates on whether females are discriminated against in consumer markets, including car retail markets (Ayres, 1991; Ayres & Siegelman, 1995; Goldberg, 1996; Morton et al., 2003) and rental housing markets (Ahmed & Hammarstedt, 2008a; Ahmed & Hammarstedt, 2008c). By contrast, even though the Equal Credit Opportunity Act (ECOA) was enacted in 1974, the issue of gender equality in credit markets, especially in the mortgage market, has attracted far less attention in the academic literature. While most of the previous studies on gender equality in credit markets focused on small-business lending (Cavalluzzo & Cavalluzzo, 1998; Coleman, 2000; Cavalluzzo et al., 2002; Blanchflower et al., 2003; Blanchard et al., 2008; Bellucci et al., 2010; Agier & Szafarz, 2012; Alesina et al., 2013), there have been very limited studies on this topic in mortgage lending literature (Black et al., 1978; Ladd, 1982; Munnell et al., 1996; Cheng et al., 2011). The topic of gender equality in the mortgage market is timely given the facts that females make up of a sizable and growing share of mortgage borrowers (Fishbein & Woodall, 2006) and that lenders have tightened regulation on loan origination after the recent financial crisis.

This study investigates whether mortgage lenders charge different loan contract rates on the basis of a borrower's gender beyond the extent to which it affects loan termination risk. Unlike the few previous studies on gender equality in mortgage lending, this study focuses on the loan origination stage instead of the loan approval stage. This focus is consistent with the lending

² Altonji & Blank (1999) and Marianne (2011) provided thorough reviews of previous research on gender equality in the labor market.

industry's shift away from a system in which lenders would deny borrowers with highest credit risk to the one in which lenders offer credit to almost everyone but at different prices (Turner & Skidmore, 1999; Ghent et al., 2014). Because of this, gender inequality may occur primarily in loan pricing instead of credit allocation.

A sample of 30-year first-lien fixed-rate subprime mortgage loans for home purchase originated in Miami-Dade County, FL from 1997 to 2006 is used to examine the influence of a borrower's gender on the loan contract rate set by the lenders. To mimic the lender's behavior at origination, the probability of a loan being prepaid or defaulted upon is estimated, using a loan hazard model based on loan performance data. The loan-level predicted default probability and prepayment probability are incorporated into a loan contract rate model to control for the degree to which a borrower's gender is associated with the probability of loan default and prepayment. This framework allows one to examine whether gender differences in the loan contract rate are attributable to gender or its impact on tendencies to default or prepay a loan.

The estimation results show that a borrower's gender is associated with the loan termination probability. Thus, it is important to control for this correlation when examining the impact of a borrower's gender on the loan contract rate. The results show that a female sole borrower is more likely to default than joint male and female co-borrowers. With this correlation controlled for in the loan contract rate model, the results indicate that female borrowers pay a higher contract rate than joint male and female co-borrowers. Specifically, the results reveal that a female sole borrower pays contract rates 13 basis points higher than joint male and female co-borrowers. Because this gender disparity in contract rate is only attributed to the gender itself, the results provide empirical evidence of gender inequality in mortgage lending.

The methodology employed in this study differs significantly from the commonly used reduced-form approach employed in prior gender studies. In a reduced-form model, the contract rate is regressed against a borrower's gender as well as a set of covariates including variables (e.g. loan characteristics, borrower characteristics) that are believed to affect the likelihood of loan termination.³ If a borrower's gender is associated with some loan termination risk factors that are unobserved, the estimates on gender in the reduced-form model would be biased. The methodology employed in this study explicitly models and accounts for the loan termination likelihood in the loan contract rate model using data on loan performance since origination. In addition, some prior related studies have very little to say about prepayment risk and its impact on the loan's contract rate while others assume prepayment risk is the predominant consideration. The current study, however, does not impose the condition that either default or prepayment is more influential than the other for loan contract rate. In fact, the prepayment hazard is examined simultaneously with the default hazard through a competing-risk hazard model, and both default probability and prepayment probability are accounted for in the contract rate model. Within the study, a distinction is made between pecuniary prepayment and non-pecuniary prepayment, based on the hypothesis that lenders would charge different risk premiums for these two types of prepayment. Furthermore, by employing a discrete-time hazard model, when modelling default hazard and prepayment hazard, the issue of left truncation and right censoring, which is common in mortgage related studies, is addressed.

The paper is organized as follows. The second section offers an overview and a discussion of previous studies on gender equality in credit markets. The third section explains the empirical model in detail. The fourth section describes the data. The fifth section explains model

³ See Cheng et al. (2011) as an example.

specifications. The sixth section presents the results. The seventh section discusses the results and final section provides the conclusion.

2.2. Literature Review

Gender equality in mortgage lending has not drawn as much attention, either politically or socially, as racial equality. One reason for the dearth of research on this topic may be the need for loan-level data on a borrower's gender. Whatever the reason, the issue of gender equality in mortgage market remains an open question.

Cheng et al. (2011) focus on loan contract rate disparity across gender groups in mortgage lending using borrower-level data from Survey of Consumer Finances (SCF). The authors find the contract rate paid by female borrowers is significantly higher than that paid by males. However, they also find that this significant disparity disappears after the search behavior variable of the borrower is controlled for in the loan contract rate determination equation. The search variable controls for whether a borrower's lender choice is based on the search for the lowest rate, or based on recommendations by others. Given the results, the authors conclude that the gender disparity in contract rate is attributable to less searching efforts by females, not discrimination by lenders.

Several studies with a focus on racial or ethnic equality in the mortgage market also shed some light on the issue of gender equality because they incorporate a borrower's gender as a control variable when examining a loan's contract rate. Overall, the results relating to gender are mixed within these studies. Cheng et al. (2014), using the same data set as Cheng et al. (2011), examine whether there is rate discrimination in lending against African Americans. The results show African Americans pay significantly higher interest rates than their white counterparts. Furthermore, quantile regression results indicate that the magnitude of the racial disparity in loan

contract rate appears to be larger for African American females relative to African American males, but the significance of this difference was not formally tested. Also using the Survey of Consumer Finances (SCF) data, Duca and Rosenthal (1994) focus on conventional fixed-rate mortgage loans, but fail to establish the existence of gender inequality in the conventional loan market. Zhang (2013) matched a proprietary loan-level data from a national bank with the Home Mortgage Disclosure Act (HMDA) data to obtain information on a borrower's gender, and concludes that single males tend to obtain a higher interest rate than joint male and female co-borrowers conditional on the loan application being approved, while single females do not.

Gender disparity in the loan approval process is also an important aspect to understand gender equality in the mortgage market. Using the Comptroller of the Currency-FDIC nationwide survey data, Black et al. (1978), fail to find any significant difference in loan denial rate between male and female applicants. Ladd (1982), studying the same issue, finds that loan applications by females in New York City are more likely to be denied than those by males. However, Ladd (1982) does not find that a borrower's gender affects the loan denial rate for loans originated in California. Using data from the Boston Federal Reserve, Munnell et al. (1996) demonstrate that males are more likely to be denied access to credit than female applicants.

While research on gender equality in the mortgage market may be sparse, gender equality has been examined in other credit markets, including small-business lending market. The results indicate that a consensus has not been reached. With the use of the National Survey of Small Business (NSSBF) data, Cavalluzzo & Cavalluzzo (1998), Blanchflower et al. (2003), and Blanchard et al. (2008) fail to establish the existence of gender inequality in small-business lending. The results show that after controlling for a large set of borrower, firm, and loan characteristics, small business loan applications by female owners are not significantly less likely to be approved,

and the interest rates charged on their loans are not significantly different from those charged on loans granted to male borrowers. However, also using the NSSBF data, Cavalluzzo et al. (2002), Coleman (2000) find empirical evidence of gender inequality in this market. Coleman (2000) demonstrates that lenders do not seem to discriminate against female owners in the form of credit allocation, but rather in the form of loan pricing and collateral requirement. Females are shown to be more likely to be asked to pledge collateral and pay higher interest rates. Cavalluzzo et al. (2002) offer evidence of gender inequality in less competitive small-business lending markets. The results demonstrate that female-owned firms experience increased loan denial rates when the measure of lender competition falls. In addition, using data from banks in Italy, both Bellucci et al. (2010) and Alesina et al. (2013) report support for the existence of gender inequity in small-business lending. Bellucci et al. (2010) conclude that female entrepreneurs are disadvantaged compared to their male counterparts in the terms of credit availability and collateral requirement, but not in the terms of loan pricing, Alesina et al. (2013) demonstrate robust evidence that female owners of small business pay more for credit than males. Agier et al. (2012), using data from a Brazilian microfinance institution find that although a gender gap does not exist in the terms of credit availability, it does exist in the terms of loan size. Results reveal that females with large business projects face a “glass ceiling” effect and experience harsher loan downsizing than men.

Overall, results from previous studies on the issue of gender inequality in credit markets show mixed results. However, when it comes to loan pricing, previous studies consistently ignore whether a borrower’s gender is associated with the loan default probability or prepayment probability. By contrast, this study explicitly accounts for the probability of loan termination in testing whether a borrower’s gender affects loan contract rate.

2.3. Model

The loan contract rate is determined by the lender at the time of loan origination. In order to accurately set the terms of a loan, the lender must carefully consider the probability of a loan being defaulted or prepaid upon by a borrower, since these decisions affect the future cash flows the lender expects to receive from the loan. Thus, the borrower's behavior with respect to loan termination is critical to the lender. Following Kau et al. (2012), the lender's loan origination behavior is modelled at origination, along with the borrower's loan termination behavior in each subsequent month since loan origination. The framework makes this paper noticeably different from previous studies in which only the lender's origination behavior was taken into account.

The lender's origination behavior is modelled differently from the borrower's termination behavior. The lender's behavior is only exhibited at the time of origination, while the borrower's termination behavior is observed at monthly intervals since loan origination. Note that a borrower's decision on whether or not to terminate a mortgage loan is only affected by the terms of the loan in a given month, and is not influenced by the lender's behavior. By contrast, in determining the contract rate of a loan at origination, the lender needs to meticulously think over the borrower's possible future behavior. The lender must consider how likely the borrower is to default or prepay based on all of the information the lender has on the loan, the borrower, and the property at loan origination. It is assumed lenders possess an accurate model of the borrower's behavior and utilize this model to predict loan default probability and prepayment probability by each borrower to approximately assess the loan termination risk and determine the contract rate. It is also assumed that in the model, the borrower's behavior can be derived from the observation of the actual borrower's behavior over time.

Within this framework, we are able to examine whether lenders take into account a borrower's gender beyond the extent to which it affects loan default probability and prepayment probability in determining the loan contract rate. It is important to note that this framework models the borrower's termination behavior with regard to both prepayment and default. These two forms of loan termination both impact the flow of proceeds from the loan, thus the value of the loan. This is an important modelling trait not found in prior mortgage literature on this and related topics.

2.3.1. The Borrower's Termination Behavior Model

The borrower's loan termination behavior is modelled using the Cox discrete-time competing-risks model (Deng et al., 2000). The discrete-time nature of the loan performance data, which is tracked monthly, is consistent with this modelling approach. From loan performance data, on the basis of the timing of the loan termination event, whether a borrower continued, prepaid, or defaulted, can be identified for each loan each month subsequent to origination (i.e., one observation for each month of each loan). This modelling approach also helps with the issue of left truncation and right censoring which are fairly common in mortgage literature. Left truncation occurs when a loan had been defaulted or prepaid upon before the start point of the observation window of loan performance, and thus is not observable. Right censoring refers to a situation where a loan has not been defaulted nor prepaid upon by the end of the observation window of loan performance. If the analysis merely classifies loans in the sample into three groups based on whether a loan was defaulted on, or prepaid at some point, or continued within the observation window of loan performance, the default probability and prepayment probability would be underestimated by this loan-level analysis (i.e. only one observation for each loan). In contrast, the discrete-time model, with the assumption that loans when unobservable follow the similar

termination pattern as loans when observable, could appropriately estimate loan termination probability.

Following the option-theoretic model of default and prepayment (Kau et al., 1995) in which default and prepayment are believed to compete with each other as a substitute, a multinomial logit model is employed at monthly intervals as in Equation (1) to model the borrower's loan termination behavior.⁴

$$\ln(p_{jt} / p_{0t}) = \delta_j \alpha_{jt} + \beta_j' \mathbf{x}_{jt} + \varepsilon_{jt} \quad j=1, 2 \quad (1)$$

where p_{0t} is the probability of a loan being current in period t ; p_{jt} is the probability of termination in period t for a loan given that this loan had not been terminated by the beginning of period t , where $j=1$ it is the probability of default at time t , and $j=2$ is the probability of prepayment at time t . Here, t refers to mortgage time.⁵ The baseline hazard rates for default and prepayment (α_{jt}) are allowed to vary across mortgage time. The vector of covariates ($\mathbf{x}_{jt}$) includes observed traits of the loan, the borrower, the property, the neighborhood, and the economic conditions. These covariates may or may not be time varying.

For each loan in each period t , the probability of a borrower defaulting on or prepaying a loan is predicted based on Equation (1). Among the time-varying covariates in $\mathbf{x}_{jt}$, it is assumed the term structure, specifically, the variation in the future mortgage interest rate would be of

⁴ Here a widely adopted method was utilized to estimate this competing risk model in which we estimated default hazard and prepay hazard separately, treating other event as censoring and assuming that one event is not informative to the other conditional on all the covariates. This was done mainly because in each hazard model we included almost all of the variables that were likely to affect both of these two events. Another reason is, for loan default and prepayment, there is no theory that could be used as the guideline to impose any parameter restrictions that cross these two hazard equations. Hence, it is not necessary to estimate default and prepay hazard models within a simultaneous equation framework, especially studies show separate models perform well for most of the data (Allison, 2010). In addition, there are other advantages of estimating default and prepay hazard models separately, including the flexibility in specifying different models for different events.

⁵ Variables in calendar time could also be expressed with mortgage time, because calendar time could be simply transformed to mortgage time with the use of loan origination month.

particular interest to the lender for the prediction.⁶ In this study, the 10-year treasury constant maturity yield is used as the benchmark for the mortgage interest rate of 30-year fixed-rate mortgage loans. Notice here, lenders are assumed to be allowed to differentiate pecuniary prepayment from non-pecuniary prepayment. Pecuniary prepayment occurs when the future 10-year yield at time t (y_t) drops below the yield at origination (y_0), while non-pecuniary prepayment occurs when the future 10-year yield at time t (y_t) is above the one at origination (y_0).⁷ The reason for allowing lenders to make a distinction between pecuniary prepayment and non-pecuniary prepayment is the former is normally driven by financial incentives from a dropping interest rate, while the latter is driven by some non-financial reasons (i.e. divorce, relocation, etc.). In addition, from the lender's perspective, in the case of non-pecuniary prepayment, lenders could reinvest the proceeds at higher interest rates, whereas in the case of pecuniary prepayment, lenders could only reinvest the proceeds at lower interest rates. Thus, it is anticipated that pecuniary prepayment is more disadvantageous to lenders than non-pecuniary prepayment, and lenders would require a higher premium to compensate the risk for pecuniary prepayment than for non-pecuniary prepayment.

The commonly used Cox, Ingersoll and Ross (CIR) term structure model was employed to predict the future 10-year yield.⁸ In the CIR term structure model, the whole term structure is

⁶ For other time-varying covariates, we either used the actual values if we were able to observe them, or extrapolated values from the known values, whichever seem more reasonable and appropriate.

⁷ In this study, we also adopted an alternative way to define pecuniary prepayment and non-pecuniary prepayment, based on the argument that borrowers would not immediately prepay when the interest rate drops just below the one at origination because of the prepayment cost as well as the option values of future prepayment and future default (Kau et al., 1995). With the alternative definition, it is deemed pecuniary prepayment occurs when the 10-year yield at time t (y_t) drops by more than 100 basis points relative to the yield at origination (y_0), and non-pecuniary prepayment occurs when the 10-year yield at time t (y_t) exceeds the yield at origination (y_0), or drops by less than 100 basis points relative to the yield at origination (y_0). The results on these two definitions are both reported in the results section, and are shown to be consistent with each other, suggesting that the way these two types of prepayment are defined does not affect the results and the conclusions. Thus, we focus on the first definition here.

⁸ Though several studies in asset pricing argue other interest rate models perform better than CIR term structure model with respect to out-of-sample prediction, those models could only be employed to forecast the mean, not the density

driven by a spot interest rate ($r(t)$). This spot interest rate is believed to follow a mean-reverting stochastic process with volatility affected by the level of the spot rate. The form for the spot interest rate is as follows:

$$dr(t) = \gamma(\theta - r(t))dt + \sigma\sqrt{r(t)}dz(t) \quad (2)$$

where the first part is the deterministic one with θ as the long-term mean of the spot interest rate and γ as the reversion rate, whereas the second part describes the stochastic movements.

Based on the estimated parameters in Equation (2),⁹ the density of future spot interest rate for any forecast interval conditional on the spot interest rate at origination $dF(r(t)|r(0))$ is forecasted, with the use of transition density of the spot interest rate implied by CIR term structure model.¹⁰ Since the change of 10-year yield is driven by the change of the spot interest rate, and the former affects loan default probability and prepayment probability, the forecasted conditional density of future spot rate $dF(r(t)|r(0))$ is utilized to predict default probability and prepayment probability as in Equations (3) - (5).

$$\hat{p}_{it}(y_0) = \int p_{it}[y_i(r(t))] dF(r(t)|r(0)) \quad (3)$$

of the spot interest rate needed here. In addition, the CIR term structure model is the standard model used in mortgage literature.

⁹ The parameters in Equation (2) were estimated with the use of 4 time series of yields with different maturities from 1987 to 2007 within the framework of the single-factor CIR term structure model. Those 4 time series are 6-month T-bill yield, 1-year Fama-Bliss bond yield, 3-year Fama-Bliss bond yield, and 5-year Fama-Bliss bond yield. Data were obtained from CRSP. The reason why we chose this estimation period (from 1987 to 2007) is many studies have found there was a shift in Federal Reserve monetary policy in the early 1980s (Duan, Simonato (1999)) and the loan data in this study ends in 2007 based on loan origination year. We used the GAUSS code offered by Jin-Chuan DUAN on his website for the estimation part, the one used by himself to yield the results in Duan, Simonato (1999). We would like to acknowledge this help from him.

¹⁰ Notice here, this study forecasts the conditional density of the future spot interest rate rather than the simple conditional mean, as the forecasted density enables the calculation of both predicted pecuniary prepayment probability and non-pecuniary prepayment probability for each loan, while forecasted mean only allows one to calculate either predicted pecuniary prepayment probability or predicted non-pecuniary prepayment probability for each loan. For the transition density of the spot interest rate, see Cox, Ingersoll, Ross (1985). Here, a normal distribution was used to closely approximate the true transition density.

$$\hat{p}_{2t}(y_0) = \int_{y_t(r(t)) \leq y_0} p_{2t}[y_t(r(t))] dF(r(t) | r(0)) \quad (4)$$

$$\hat{p}_{3t}(y_0) = \int_{y_t(r(t)) > y_0} p_{3t}[y_t(r(t))] dF(r(t) | r(0)) \quad (5)$$

Here, $\hat{p}_{1t}(y_0)$ is the predicted default probability in period t seen from loan origination given that this loan had been continued by the beginning of period t , $\hat{p}_{2t}(y_0)$ is the predicted pecuniary prepayment probability in period t , and $\hat{p}_{3t}(y_0)$ is the predicted non-pecuniary prepayment probability in period t . Here, the integrated expectations are numerically approximated through a discretization approach in which the spot interest rate domain was divided into numerous but finite intervals.

The predicted probability of each loan's termination event in any particular period t , from Equations (3) - (5), is aggregated over a 10-year span to arrive at the total predicted probability of each event $\hat{P}_k$ ($k=1, 2, 3$) seen from origination as in Equation (6). This reflects the lender's concern regarding the total predicted probabilities at origination rather than in each time period.¹¹

$$\hat{P}_k = \sum_{t=1}^T \left(1 + \frac{y_0}{12}\right)^{-t} \hat{p}_{kt} \prod_{s=1}^{t-1} \left(1 - \sum_{k=1}^3 \hat{p}_{ks}\right) \quad k = 1, 2, 3, \quad (6)$$

Here, $\hat{p}_{kt}$ is the predicted probability of event k in period t given that the loan had survived by the beginning of period t with the probability as $\prod_{s=1}^{t-1} \left(1 - \sum_{k=1}^3 \hat{p}_{ks}\right)$. Hence, $\hat{p}_{kt} \prod_{s=1}^{t-1} \left(1 - \sum_{k=1}^3 \hat{p}_{ks}\right)$ is the unconditional predicted probability of event k in period t . These probabilities are discounted by the 10-year yield at origination (y_0) with the assumption that lenders are more concerned with loan termination at earlier stages of the loan. The total predicted probability of each event ($\hat{P}_k$) is the

¹¹ Notice here, a capital P is used to distinguish total loan termination probabilities from time-specific loan termination probabilities. The subscript k tells the type of the event, 1 for default, 2 for pecuniary prepayment, and 3 for non-pecuniary prepayment.

summation of time-specific discounted unconditional predicted probability of event k over a 10-year window.

2.3.2. *The Lender's Origination Behavior Model*

Since these three total predicted probabilities ($\hat{P}_k$) derived from the borrower's loan termination behavior model appropriately represent the expectations of the lender at the time of loan origination, they are incorporated into a loan contract rate determination equation (Equation (7)) that models the lender's loan origination behavior.

$$C_0 = \alpha_0 y_0 + \beta_1 \hat{P}_1 + \beta_2 \hat{P}_2 + \beta_3 \hat{P}_3 + \gamma' \mathbf{z} + \varepsilon \quad (7)$$

In Equation (7), in addition to the three generated regressors (total predicted default possibility and prepayment possibilities), the 10-year yield at origination (y_0) is incorporated, and this rate is believed to be a fundamental factor in determining a loan's contract rate. Additionally, a full set of covariates at loan origination are also included in $\mathbf{z}$, including the characteristics of the loan, the borrower, the property, and the neighborhood. The estimates from Equation (7) will allow us to examine whether a borrower's gender played a role determining the loan contract rate, in addition to the loan contract rate being rationally determined by the current yield and various risk premiums that reflect the probability of the borrower defaulting or prepaying the loan.

2.4. Data

The data in this study consist of 30-year fixed-rate mortgage loans serviced by GMAC Residential Capital Company, LLC (GMAC ResCap). GMAC ResCap was a real estate finance company that specialized in servicing subprime residential mortgage loans. Loans in the data were originated by different lenders, and then bundled into residential mortgage-backed securities and

sold in the secondary mortgage market. From the loan origination data, detailed information on loan characteristics and borrower financial characteristics was obtained. In addition to loan origination data, information on monthly performance of each loan was obtained from GMAC ResCap servicing records. The loan origination data was matched to the loan performance data through a unique loan identification number created by this servicer. The loan performance data include the current balance of the loan, as well as prepayment and delinquent status of each loan on a monthly basis. In this study, default is defined as the occurrence of a borrower being 90-days delinquent, and that occurrence eventually leads to a foreclosure.

The loan origination and servicing information does not identify a borrower's gender. However, information on a borrower's gender is available in mortgage documents recorded by local governments. To make use of this source of gender information, the loan data is restricted to residential home purchase mortgage loans with underlying properties located in Miami-Dade County, FL to match with a source of readily available property transactions. The loan origination data was matched to property sale data offered by the Office of the Property Appraiser in Miami-Dade County to identify the property securing each loan in the sample based on a series of transaction characteristics. Those characteristics include the value of the underlying property, property sale month (loan origination month), property type, and zip code at sale, resulting in a matched sample of loan-property sales.¹² For each of the property transactions matched to a loan, a deed document was obtained from the records system. The deed document provided detailed information on the grantor(s) and grantee(s) including their name, gender, and marital status. With

¹² Each mortgage loan was matched to property sales in the pool with replacements requiring that the gap between the appraised value of the property in the loan data and the transaction price of the property in property sale data is the minimum one in the pool. In some cases, one mortgage loan has multiple property sale matches with the same minimum gap. However, information on deed document and mortgage document linked to each property sale could be used to identify the correct unique match for each mortgage loan, and the procedure was described below.

the use of their name and property sale date, the corresponding mortgage document was manually searched in the records system. Within the mortgage document, the number of borrowers who signed their name, gender, and marital status could be accurately identified. In addition, note date and original loan amount were also included in the mortgage document. Hence, we were able to verify the accuracy of each loan-property sale match by requiring that the note date and original loan amount in the mortgage document be identical to those in the loan origination data.

By identifying the unique property transaction which secures the loan, the property's location can be used to gather neighborhood traits for the loan. A property's neighborhood is defined as its location based on the 1990 census tract boundaries. Using this information, neighborhood characteristics such as housing occupancy rate, poverty rate, average household income, and the proportion of African Americans can be obtained. Time-varying variables of neighborhood characteristics are generated using a linear time-trend between the decennial census survey data from 1990 to 2000 and 2000 to 2010 normalized to the 1990 census tract boundaries.

The change in house prices, as well as the variation in house prices, within a census tract (neighborhood) is calculated using property transaction information.¹³ A median housing price index is generated for each month of the analysis by creating a three-year window of sales, eighteen months before and eighteen months after, for each census tract and calculating the median price.¹⁴ The result is a unique monthly median house price index for each census tract. This index is used to measure the changes in house prices over the life of the loan relative to origination. The standard deviation of house prices is also calculated for each 3-year window and is utilized to measure the heterogeneity in housing sale price in a neighborhood. In addition, for each loan at the time of

¹³ The property transaction data are sales over the 1990 to 2013 period in Miami-Dade County, FL.

¹⁴ Notice here, we chose the median house sale price instead of average house price in order to prevent any extreme house sale prices in a neighborhood from affecting the measurement of overall house price level.

origination, the neighborhood-level recent housing price appreciation rate is calculated using the growth rate of the calculated median housing sale price within a pre-origination window.¹⁵

In the data set, there were initially 4,790 loans that were originated in Miami-Dade County, FL. Of these 4,790 loans, 3,419 loans were correctly and uniquely matched to property sales. Loans with missing values on the loan, borrower, property, or neighborhood characteristics, or loans without loan performance data were deleted. In addition, the sample was restricted to loans originated from 1997 to 2006 simply because there were few subprime loans originated before 1997 or after 2006 in this data set. Furthermore, we only included loans with an underlying property sold through a warranty deed. The final sample consists of 2,206 observations of 30-year first-lien fixed-rate residential primary mortgage loans for home purchase. They were originated from Jan. 1997 to Dec. 2006. Monthly performance of these loans was observed from Jan. 2000 to Oct. 2010, a period that covers the recent financial crisis.

Tables 2.1 and 2.2 contain summary statistics on the loans in the final sample. Table 2.1 provides a brief description of the characteristics of the loan, the borrower, the underlying property, and the neighborhood at the time of origination for the pooled sample. Loans in this sample cover most of the census tracts in Miami-Dade County, FL, which allows for great variation on the neighborhood characteristics. Because loans in the sample are subprime loans, the feature of high credit risk of borrowers is demonstrated by the average original LTV ratio and the proportion of borrowers who failed to provide full income documentation. The average original LTV ratio is around 85%, and 154 loans (6.98%) have their original LTV ratio exceeding 100%. In the final sample, only 47.96% of the borrowers provide full income documentation. The proportion of loans

¹⁵ Recent housing price appreciation rate at origination is defined as the ratio of the median housing sale price in a neighborhood in a three-year period prior to the month of loan origination to the median housing sale price in the same neighborhood in another three-year period prior to the three-year pre-origination period, then minus 1.

that are encumbered by a prepayment penalty is approximately 36%. The spread between the contract rate and 10-year treasury constant maturity yield indicates a high risk premium. The average contract rate is 8.04%, approximately 308 basis points higher than the 10-year treasury constant maturity yield. Of the 2,206 loans, 9.43% (208 loans) in the sample were defaulted upon and ended in foreclosure, while approximately 80% of the loans were prepaid during the study period.

Table 2.2 provides descriptive statistics of the loan sample by the gender of a borrower(s). Borrowers are broken into three gender-based categories and are defined as: a male sole borrower, a female sole borrower, and one male and one female co-borrowers. More than half of the loans (56.71%) in the final sample were originated jointly by male and female co-borrowers. The number of the loans originated by female sole borrowers is quite close to that by male sole borrowers. This confirms the fact that females make up a sizeable share of mortgage borrowers (Fishbein & Woodall, 2006).

The descriptive statistics of the three gender-based categories, found in Table 2.2, may provide some information on gender-based steering in the mortgage market. Gender-based steering occurs when borrowers are steered towards risky and high-cost subprime loans simply because of their gender. If steering exists on the basis of a borrower's gender, it would be anticipated that the quality of female borrowers on average is higher than that of borrowers in other gender groups in the subprime sector. Among the three gender groups, joint male and female co-borrowers, appear to have loans with the lowest original LTV ratio and the largest loan size. Differences in FICO score among the three gender groups appear to be small on average. Also a female sole borrower appears to be less likely to offer full income documentation than borrowers in other groups. Overall, these descriptive statistics do not appear to indicate that in the subprime sector, female borrowers

tend to have higher credit quality than borrowers in other gender groups. In addition, among the three gender groups, a female sole borrower appears to default on average more frequently than borrowers in other groups in the sample. The observed default rate by a female sole borrower is about 14.64%.

2.5. Model Specifications

2.5.1. *The Borrower's Termination Behavior Model*

In the borrower's loan termination behavior model, the baseline hazard rates α_{jt} , together with all other covariates in $\mathbf{x}_{jt}$, are used to model the borrower's decision to terminate a loan. Given the option-theoretic model of "financial" termination (Kau et al., 1995), financial motivations for default and prepayment are not the same; therefore, different covariates are included in the two hazard equations. In the default model, a scaled standard Default Assumption schedule (SDA) is used, while for prepayment, mortgage year fixed effects are used to allow for more flexibility in baseline hazards.¹⁶

To measure the market interest rate change at time t , the gap between the 10-year treasury constant maturity yield at loan origination and the 10-year yield at time t lagged by 2 periods ($y_0 - y_{t-2}$) is measured.¹⁷ As Kau et al. (1995) note, another prominent time-varying covariate is the house price change at time t seen from loan origination.¹⁸ Recall that to measure the change in house prices, a census tract-level median house price index is generated from the

¹⁶ The traditional Public Securities (PSA) schedule is not used because previous studies argued this schedule did not describe the pattern of actual prepayments well, for more details, see Kau et al. (2004).

¹⁷ Notice that, the 10-year treasury constant maturity yield at time t lagged by 2 periods is used as the yield at time t for every mortgage time t except for the first mortgage month and the second mortgage month. For these two months, the 10-year yield at loan origination (at time 0) is used. A yield at time t lagged by 2 periods is used because in practice there is usually a gap between a borrower's decision and actual termination, and borrowers typically rely on past information to make their decisions.

¹⁸ See Kau et al. (1995).

property transaction data base in Miami-Dade County, FL with the assumption that the value of a house changes at the same rate of the median house price index in the census tract where the house is located. For each month, this index is calculated based on the inflation-adjusted median housing sale price over a three-year window around that month, eighteen months before and eighteen months after that given month. All prices are defined in 2009 dollars.¹⁹ In order to reflect the house price change at mortgage time t relative to house prices at origination, the ratio of the median house price index level at time t to the median house price index level at origination is found. This relative house price measure at time t is denoted as RHP_t . In order to control for any correlation between market interest rate change and house price change, an interaction term of the market interest rate change ($y_0 - y_{t-2}$) and relative house price at time t (RHP_t) is also included in the loan hazard model.

Other covariates on loan characteristics, borrower characteristics, and property characteristics are also incorporated, including the loan contract rate spread at origination ($C_0 - y_0$), original LTV, FICO score of the borrower, original loan amount, loan origination season fixed effects, and dichotomous (0,1) indicators for whether a borrower provided full income documentation, whether the loan is encumbered by a prepayment penalty at time t , whether the underlying property is occupied by the owner, and whether the property is a single family detached house or a condo.²⁰ Also, several time-varying variables are included to measure the characteristics and evolution of the neighborhood of the underlying property, including the heterogeneity in the housing sale price, occupancy rate, average household income, poverty rate, and the proportion of

¹⁹ Prices are adjusted by a GDP per capita deflator.

²⁰ We chose contract rate spread instead of contract rate itself because the benchmark interest rate (10-year yield) varied considerably within our study period, and this spread allows us to make comparison within mortgage vintage.

African Americans at time t .²¹ The variables representing a borrower's gender are also included throughout the analysis. Furthermore, in the default equation, prepayment penalty variable at time t is excluded, and in the prepayment equation, the original LTV ratio, the income documentation status variable, and the variable of neighborhood-level heterogeneity in housing price at time t are excluded, because these variables are not believed to directly affect the corresponding hazard respectively.²²

Among those covariates discussed above, one variable needs more attention, the loan contract rate spread at origination ($C_0 - y_0$). This variable will be endogenous if the contract rate (C_0) is endogenous, because it is simply a linear function of the contract rate. It might be possible that the contract rate is endogenous, because as a lender believes a borrower is more likely to default or prepay, the lender would charge a higher contract rate; meanwhile being charged a higher contract rate, a borrower may be more likely to terminate the loan through either default or prepayment. In order to solve this issue and to yield unbiased and consistent estimates of the multinomial logit model, a control function (CF) method is employed. With the CF method, a reduced-form estimation of the contract rate is conducted where the contract rate is regressed against all of the exogenous variables in the system. The residual from this reduced-form estimation is then included in the multinomial logit model as an additional covariate. The estimated coefficient on the residual thus shows whether the issue of endogeneity exists.

²¹ Heterogeneity in housing sale price at time t was measured by the standard deviation of housing sale price within a three-year window prior to time t in a neighborhood. Average household income in a neighborhood at time t was measured by the ratio of the average household income in a neighborhood at time t to the average household income in Miami-Dade County, FL at time t .

²² We tested whether those excluded variables affected the corresponding hazard respectively. None of the estimated coefficients of those variables are statistically significant.

2.5.2. The Lender's Origination Behavior Model

In the lender's origination behavior model, the aggregated predicted loan default probability and prepayment probabilities ($\hat{P}_k$) are entered as generated variables. Recall that $\hat{P}_k$ is a function of the contract rate spread ($C_0 - y_0$), since the contract rate spread is included in the multinomial logit model and the multinomial logit model is used for the calculation of $\hat{P}_k$. Therefore, these generated variables ($\hat{P}_k$) are endogenous in the final contract rate determination equation. This issue is solved by a set of generated IVs ($\tilde{P}_k$), each as a valid IV for the corresponding generated variable ($\hat{P}_k$). To generate the set of IVs, the procedure to calculate $\hat{P}_k$ is used to calculate $\tilde{P}_k$, but with the actual contract rate spread ($C_0 - y_0$) being replaced with a predicted contract rate spread ($\hat{C}_0 - y_0$). The predicted contract rate spread is the difference between the predicted contract rate at loan origination ($\hat{C}_0$) and the actual 10-year yield at origination (y_0). The predicted contract rate at loan origination ($\hat{C}_0$) was obtained by the contract-rate reduced-form estimation. Because the predicted contract rate is a function of all the exogenous variables at origination, the generated IVs ($\tilde{P}_k$) are exogenous and serve as valid IVs for the three generated regressors in the loan contract rate determination equation. Because the final contract rate determination equation is a linear model, 2SLS is employed to implement the estimation.

In this model, in addition to the total predicted loan termination probabilities, the 10-year yield at origination (y_0) was also included as well as a full set of covariates at loan origination in $\mathbf{z}$ that theory suggests should affect loan contract rate. Covariates in $\mathbf{z}$ include original LTV, FICO score of the borrower, original loan amount at origination, loan origination season fixed effects, prepayment penalty fixed effects, a dichotomous variable for whether a borrower provided full income documentation at loan origination, underlying property type fixed effects, property occupancy status fixed effects, and a list of neighborhood characteristics at loan origination

including recent house price appreciation rate, heterogeneity in housing sale price, occupancy rate, average household income, poverty rate, and the proportion of African Americans.²³ Additionally, a trend term in calendar time was also included. Table A.1 in the Appendices lists all of the variables used in this study and provides a description of each variable in detail.

In the two models some covariates on the characteristics of the loan take a nonlinear function form based on prior research or some theoretical reasons. The original LTV ratio is transformed into categorical variables: loans with LTV ratio less than or equal to 75%, loans with LTV ratio greater than 75% but less than or equal to 80%, loans with LTV ratio greater than 80% but less than or equal to 90%, loans with LTV ratio greater than 90% but less than or equal to 100%, and loans with LTV ratio exceeding 100%. Furthermore, FICO score is entered as a continuous linear spline function with a knot point at 700 based on the assumption that an additional increase in FICO score has little effect on loan termination probabilities/loan contract rate when FICO score is above 700.²⁴ Finally, following prior studies, a quadratic function form of the original loan amount is also utilized to allow a non-linear relationship between original loan size and loan termination probabilities/loan contract rate.

²³ Recent house price appreciation rate at origination in a neighborhood was described by the growth rate of the median housing sale price prior to loan origination. Specifically, it was defined as the ratio of the median housing sale price in a neighborhood in a three-year period prior to month of loan origination to the median housing sale price in the same neighborhood in another three-year period prior to the three-year pre-origination period just mentioned, then minus 1. Heterogeneity in housing sale price at origination was defined as the standard deviation of the housing sale price over a three-year period prior to the month of loan origination. Average household income at origination in a neighborhood was defined as the ratio of the average household income at origination in that neighborhood to the average household income in Miami-Dade County, FL.

²⁴ The FICO linear spline function was specified as follows: $FICO_{(FICO \leq 700)} = \text{minimum}(FICO, 700)$; and $FICO_{(FICO > 700)} = \text{maximum}(FICO, 700) - 700$. Therefore, coefficient on $FICO_{(FICO \leq 700)}$ measures the effects of FICO score on dependent variable when $FICO \leq 700$; while coefficient on $FICO_{(FICO > 700)}$ measures the effects of FICO score when $FICO > 700$. We tested whether the results are robust to the specification of the FICO score knot point by conducting the same analysis with a knot point at 720 or 750, and results are robust.

2.6. Results

2.6.1. *The Borrower's Termination Behavior Model*

Table 2.3 contains the estimation results from the multinomial logit model of the borrower's termination behavior. The estimates from the default model, model (1), indicate that the default decision is driven by the financial considerations. The coefficient estimate on the relative housing price at time t (RHP_t) is negative and significant. This is consistent with option theory in that house price change is a significant factor affecting a borrower's decision to default. As house price increases, borrowers would have less incentive to default. In addition, the results also indicate that borrowers with a higher original LTV ratio, a lower FICO score, a larger original loan size, or limited income documentation, tend to be more likely to default. Note that the effect of original loan size on default probability diminishes as the original loan size increases. Furthermore, the results show that a borrower using the loan to purchase a house as the primary residence is less likely to default compared to a borrower using the loan to make an investment on the house non-owner occupied. Homebuyers of condos are more likely to default than homebuyers of single-family detached houses. Finally, the results on the neighborhood characteristics demonstrate that loans originated in a lower income neighborhood (measured as the census tract average household income relative to the county) or loans originated in a neighborhood with a higher proportion of African Americans are associated with a higher probability of default. The results are consistent with most of the prior studies.

Turning the attention to the results on the gender variables, the results demonstrate that after controlling for all of the covariates, a female sole borrower tends to be more likely to default relative to joint male and female co-borrowers. The coefficient on this gender category variable is significant at a 5% level.

The results on the prepayment hazard, model (2) in Table 2.3, are also consistent with expectations. The option-theoretic model of default and prepayment suggests a borrower's decision to prepay is induced by both the house price change and the market interest rate change (Kau et al., 1995). Specifically, in a market with falling interest rates, mortgage loan borrowers would have great incentives to exercise the prepayment option and take advantages of lower interest rates. In a period of increasing house prices, borrowers may prepay their loans in order to capitalize on house price appreciation. In addition, as prepayment and default are competing risks, an increase in house price would reduce default probability, and thus raise the prepayment probability. The second channel is an indirect effect. Both of the two channels predict that an increase in house price is associated with an increase in prepayment probability. The estimation results on the prepayment hazard are consistent with these expectations, as the coefficients of market interest rate change at time t ($y_0 - y_{t-2}$) and relative house price at time t (RHP_t) variables are both shown to be positive and statistically significant.

A borrower's prepayment probability is also affected by factors other than changes in house prices and interest rates. The contract rate spread at origination ($C_0 - y_0$) appears to be positively associated with the probability of loan prepayment. If the contract rate spread represents a risk premium, the result would indicate borrowers with a larger risk premium would be more likely to prepay. Another possible explanation could be the effects of points on the spread. Higher level of points paid would result in a lower spread. As theory suggests, borrowers, anticipating a higher likelihood of future prepayment, self-select less points, thus the presence of points is predicted to negatively affect the prepayment probability. However, in the data set, points paid by a borrower at origination could not be observed, thus the effect of points on prepayment could not be directly tested. The results also indicate that a borrower with a lower FICO score, a smaller loan size, or a

prepayment penalty tends to be less likely to prepay. The results also indicate that a borrower living in a condo versus a single-family detached house, or a borrower living in a poorer neighborhood tends to be less likely to prepay, *ceteris paribus*. After controlling for all of those covariates, the results in the prepayment hazard model do not find any evidence that a borrower's gender has a significant impact on prepayment probability.

In addition, note that the residual generated from the contract-rate reduced-form regression is shown to be insignificant in both the default hazard and prepayment hazard estimations, indicating that both the model for the borrower's termination behavior and the reduced-form model for the lender's origination behavior are well specified.²⁵

2.6.2. The Lender's Origination Behavior Model

Table 2.4 reports the results on the contract rate equation for two specifications. The specifications differ in terms of how the pecuniary prepayment and non-pecuniary prepayment are defined. In model (1), pecuniary prepayment is defined to occur when the 10-year yield at time t (y_t) drops below the yield at origination (y_0), while non-pecuniary prepayment occurs when the 10-year yield at time t exceeds the yield at origination. In model (2), an alternative definition for these two types of prepayment is adopted, mainly based on the argument that borrowers would not immediately prepay when the interest rate drops just below the rate at origination because of the cost of prepayment as well as the option values of future prepayment and future default (Kau et al., 1995). Thus, in specification (2), pecuniary prepayment is defined to occur when the 10-year yield at time t (y_t) drops by more than 100 basis points relative to the yield at origination (y_0), and non-pecuniary prepayment occurs when the 10-year yield at time t (y_t) exceeds the yield at

²⁵ The contract-rate reduced-form estimates were reported in Table A.2 in the Appendices.

origination (y_0), or drops by less than 100 basis points. The definition of these two types of prepayment does not affect the results nor the conclusions. Recall that the predicted default probability and prepayment probabilities are entered into the contract rate determination equation as generated regressors. Therefore, the reported *p-values* in Table 2.4 are calculated using the corrected standard errors.²⁶

The results in Table 2.4 indicate that the predicted default probability, pecuniary prepayment probability, and non-pecuniary prepayment probability all have positive impacts on the contract rate, indicating lenders charge positive risk premiums for both default risk and prepayment risk. A 10-percentage-point increase in the default probability would lead to an increase in the contract rate by 12 to 14 basis points. The increase in the contract rate from a 10-percentage-point increase in the probability of pecuniary prepayment ranges from 22 to 27 basis points, while that from a 10-percentage-point increase in the probability of non-pecuniary prepayment ranges from 14 to 17 basis points. The results indicate that pecuniary prepayment is more disadvantageous to lenders than non-pecuniary prepayment. In addition, the significant and positive coefficients of these three generated variables confirm the importance of incorporating them into the loan contract rate determination equation.

Many of the estimates from the contract rate equation are consistent with prior expectations. The estimate on this 10-year yield at origination variable is positive and statistically significant. Interestingly, the coefficient is less than one. The LTV ratio at origination is positively associated with the contract rate, confirming the proportion of the equity a borrower has on the house at loan origination is a key factor for loan pricing. A borrower's FICO score has a negative association with loan contract rate, but we fail to reject that this impact varies for FICO scores above 700.

²⁶ Standard errors were corrected using the standard way demonstrated in Appendix 6A of Wooldridge (2010).

Borrowers without full income documentation tend to be charged higher loan contract rates. In addition, lenders tend to charge a lower contract rate to a borrower who used the loan to purchase a house as the primary residence versus a borrower who used the loan to make an investment on the house.

The results also indicate that larger loan size reduces the loan contract rate, but at a decreasing rate. The finding on loan size is consistent with most of the prior studies and there are two possible explanations for it. First, loan size is believed to be positively correlated with a borrower's income and wealth, and the latter is anticipated to be negatively associated with the loan contract rate. Therefore, from this point of view, loan size is expected to be negatively correlated with loan contract rate. Second, given that the loan-serving fees do not depend on loan size, as loan size increases, loan-servicing fees per unit of original loan amount decrease. As loan-serving fees have been incorporated in the contract rate, we would also expect a negative relationship between the contract rate and the original loan amount. The results also indicate that prepayment penalty is associated with the contract rate. At first glance, it seems that the presence of prepayment penalty is positively associated with the contract rate, as the coefficients on the four prepayment penalty category variables are all significantly positive. This finding is somewhat puzzling given that prepayment penalty transfers parts of the prepayment risk from the lender to the borrower, and thus should reduce the contract rate. However, a close examination reveals that the longer a loan is encumbered by a prepayment penalty, the lower the contract rate is, indicating the prepayment penalty itself does reduce the loan price. One possible explanation for the positive and significant coefficients on the four prepayment category variables is borrowers selecting a prepayment penalty tend to be riskier, other things equal.²⁷ Furthermore, the results on the

²⁷ A prior study by Mayer et al. (2013) provided evidence for this explanation.

neighborhood characteristics show that lenders take into account the neighborhood-level income and recent house price appreciation rate for their decisions on the loan contract rate. As indicated in Table 2.4, borrowers living in affluent neighborhoods, or in neighborhoods experiencing house price appreciation, appear to receive lower rates.

Turning attention to the results on the gender category variables which are of our primary concern, female borrowers are shown to pay more for their mortgage loans. A female sole borrower is shown to face a contract rate 13 basis points higher than joint male and female co-borrowers, *ceteris paribus*. As this gender disparity in the contract rate is not attributable to differences in default probability or prepayment probability, but rather due to the gender itself, this result provides evidence of gender inequality in mortgage lending.

The models presented in Table 2.4 fail to disentangle the effects of gender and race on a loan's contract rate. The coefficient on gender may capture the impact of race if female borrowers in the study are more likely to be African Americans. To address this possibility, in the absence of individual race trait, an interaction term of a borrower's gender and race is included in the loan contract rate determination model. As data on a borrower's race is not available in this study, a variable indicating whether a borrower lives in a predominantly African American neighborhood is used to proxy a borrower's race, given the assumption that a borrower living in a predominantly African American neighborhood might be more likely to be an African American.²⁸ The results presented in Table 2.5 remain virtually unchanged from the previous models. The coefficients on the interaction variables are not found to be statistically significant and the coefficient on the

²⁸ It is defined that a neighborhood is a predominantly African American one if more than 50% of the residents in that neighborhood are African Americans. We also used alternative ways to define a predominantly African American neighborhood. With alternative definitions, a neighborhood is defined as a predominantly African American one if more than 60% (or more than 70%) of the residents in that neighborhood are African Americans. The results are consistent with those alternative definitions.

female variable remains positive and significant. These results indicate that it is a borrower's gender rather than race that has a significant impact on the contract rate.

2.6.3. Robustness Tests

The gender disparity in the contract rate found in Table 2.4 is robust to alternative specifications and nonparametric matching tests.

To explore the robustness of these results, we change the lens of our analysis to a nonparametric approach. Gender equality studies, especially focusing on the wage gap in the labor market, have utilized parametric and nonparametric approaches. The traditional regression-based wage gap decomposition approach developed by Blinder and Oaxaca in 1973 has been widely used for 40 years (Altonji & Blank, 1999; Marianne, 2011). However, several recent studies have developed nonparametric matching approaches to explore this issue (Ñopo, 2007; Frölich, 2007; Black et al., 2008). The nonparametric matching approach developed below is similar to those used in the labor economics literature.

The nonparametric matching approach is solely based on the hypothesis that a borrower with the same level of loan termination risk should pay the same contract rate regardless of the gender. In this study, as the level of loan termination risk is fully measured by the predicted probability of each event including default, pecuniary prepayment, and non-pecuniary prepayment, matching is conducted based on the three predicted probabilities as well as loan origination

month.²⁹ Nearest 1-to-1 matching with replacements and with calipers is performed to match loans of a female sole borrower to loans of a male sole borrower.³⁰

The matching algorithm is summarized as follows. First, for each loan of a female sole borrower in the sample, all the loans of a male sole borrower that satisfy the calipers are selected as potential “matches”. A caliper is used to impose a tolerance level for the maximum distance on a matching criterion between a loan of a female sole borrower and a loan of a male sole borrower.³¹ Then, for each loan of a female sole borrower, within the resulting set of potential “matches” by the first step, the loan with the smallest total difference in loan termination probability is selected as the sole and best match, and is included in the matching sample.³² Finally, in order to check whether the female group and the male group in the resulting matching sample have the same distributions of loan termination risk and baseline mortgage interest rate at origination, four balance diagnostic variables are used to diagnose whether the matching sample is well balanced.³³ They are the predicted probability of each event (including default, pecuniary prepayment, and non-pecuniary prepayment) and the 10-year treasury constant maturity yield at loan origination (y_0). Based on the assumption that the contract rate of a loan is wholly determined by the level of

²⁹ The origination month is used as a matching criterion because a loan’s contract rate is affected by the baseline mortgage interest rate in the origination month, and the baseline mortgage interest rate varies with time.

³⁰ Only loans of a male sole borrowers and loans of a female sole borrowers are included for matching because in such way the effects of number of borrowers on the contract rate could be eliminated. In addition, in this study, the number of loans of a sole borrower is sufficient to conduct nearest 1-to-1 matching.

³¹ The matching criteria include the predicted probability of each event (including default, pecuniary prepayment, and non-pecuniary prepayment), loan origination month, and the geographic location of the property securing the loan.

³² The total difference in loan termination probability is defined as the quadratic mean of the differences in the predicted default probability ($k=1$), the predicted pecuniary prepayment probability ($k=2$), and the predicted non-pecuniary prepayment probability ($k=3$) between a female sole borrower and a male sole borrower. Specifically, it is defined as $|\hat{p}^{female} - \hat{p}^{male}| = \sqrt{\sum_{k=1}^3 (\hat{p}_k^{female} - \hat{p}_k^{male})^2 / 3}$, where $\hat{p}_k^{female}$ is the predicted probability of event k for a female

borrower, and $\hat{p}_k^{male}$ is the predicted probability of event k for a male borrower.

³³ The standardized difference of the mean and the ratio of the variance of each balance diagnostic variable between the female group and the male group are used to measure whether each matching sample is well balanced following the guidelines in Rubin (2001).

loan termination risk and the baseline mortgage interest rate at loan origination in a gender-blind world, a test is conducted to examine whether the two gender groups have the same distribution of the contract rate.

The results by the nonparametric matching approach are robust to various specifications of the calipers, with each caliper resulting in a different matching sample as showed in Table 2.6.³⁴ The matching results demonstrate that in each matching sample, the female group and the male group have the same distribution of the baseline mortgage interest rate at origination measured by the 10-year treasury constant maturity yield, as well as the same distributions of the level of loan termination risk measured by the three predicted probabilities, however, the distribution of the loan contract rate differs between the two gender groups. Specifically, the contract rate of the female group is shown to be significantly higher than that of the male group; a result consistent with the parametric results.

2.7. Discussion

Given the main empirical findings that gender disparity in loan pricing exists and the disparity cannot be explained by differences in loan default probability or prepayment probability, we provide several explanations for the possible sources of the disparity, although our data limits our ability to identify the specific source.

³⁴ Specifications of the calipers vary by the maximum difference acceptable for matching on the predicted probability of each event k ($k=1, 2, 3$) between a loan of a female sole borrower and a loan of a male sole borrower, as well as by whether a caliper on the distance of the geographic location of the property securing a loan is applied. In Table 2.6, for matching sample 1 and 2, it is required the difference in the predicted probability of each event k between a loan of a female sole borrower and a loan of a male sole borrower be less than or equal to 5%, while for matching sample 3 and 4, the difference in the predicted probability of each event k should be less than or equal to 2.5%. Meanwhile, for matching sample 1 and 3, the location distance caliper is not applied, while for matching sample 2 and 4, the location distance caliper is applied, with a maximum distance of 7.5 miles. Location distance calipers of 10 miles and 12.5 miles are also applied respectively. The results are consistent and are omitted in Table 2.6, but are available upon request. Additionally, a caliper on loan origination month distance is applied for all of the matching samples, with a maximum distance of 18 months.

One limitation this study faces is number of points paid by a borrower at loan origination is not available in the data. Points are negatively correlated with a loan's contract rate. Thus it might be possible that higher contract rates paid by female borrowers might be due to less points paid by female borrowers on average. However, this explanation is unlikely to be the main reason for the gender disparity found in this study. The results of the loan hazard model in Table 2.3 do not indicate female borrowers are more likely to prepay. Borrowers with lower points should be more likely to prepay. In addition, while there might be a reason to believe that female borrowers might have less wealth than comparable male borrowers, and hence they might pay less points, a prior study found that borrowers in disadvantaged groups tend to be charged more for loan closing cost (Woodward, 2008). Finally, there is no hard evidence that female borrowers pay significantly less points than comparable male borrowers.

A second explanation, in line with Cheng et al. (2011), for a rate differential could be due to differences in the behavior of borrowers across gender groups. Loan contract rates are determined through negotiation by a borrower and a lender. If there are differences across gender groups in the ability of borrowers to search for the lowest rate, to effectively compare across multiple loan offerings with different sets of loan terms, or to negotiate with lenders, gender disparity in loan pricing may be due to the differences in the behavior of the borrowers. Some previous studies have found that males trade more aggressively (Barber & Odean, 2001), and have stronger bargaining power than females (Niederle & Vesterlund, 2007; Eckel et al., 2008; Vandergrift & Yavas, 2009). Thus, this source on the borrower's side could be possible. This source of variation could be detected by including variables in the loan contract rate determination equation that describe a borrower's loan searching efforts, bargaining power, and knowledge on

the mortgage market. However, when all is said and done, women are still charged a different rate, because of a trait associated with gender.

Finally, the rate disparity could simply be discrimination. Given the findings by previous studies that gender discrimination existed in small-business lending in the form of credit allocation, loan pricing, or collateral requirements (Coleman, 2000; Cavalluzzo et al., 2002; Bellucci et al., 2010; Agier et al., 2012; Alesina et al., 2013), it would not be surprising if it also exists in the subprime mortgage market, the market in which lenders face less stiff scrutiny by government agencies.

2.8. Conclusion

This study attempts to measure the impact of a borrower's gender on the loan contract rate beyond the extent to which it affects loan default probability and prepayment probability. The traditional approach of estimating a contract rate model with a gender fixed effect variable and the absence of the probability of termination variables fails to recognize that gender could proxy for termination behavior. To address this problem, predicted loan default and prepayment probabilities implied by the loan hazard evaluation are incorporated into the loan contract rate determination model along with gender traits.

Based on a sample of 30-year fixed-rate home-purchase subprime loans originated from 1997 to 2006 in Miami-Dade County, FL, the results provide empirical evidence of gender inequality in the subprime mortgage lending. Results on the loan hazard evaluation demonstrate that loan default probability is associated with a borrower's gender group. A female sole borrower appears to be more likely to default than borrowers in other gender groups. The results on the loan contract rate determination model provide evidence of adverse pricing against female borrowers

in the subprime mortgage market, and this adverse pricing is solely due to gender, per se. The results indicate a female sole borrower appears to pay 13 basis points more for the contract rate than joint male and female co-borrowers (the base gender group in this study), other things equal.

In the analysis, we also provided explanations for the possible sources of the observed gender disparity in loan contract rate. Those sources include discrimination against females by lenders in the mortgage market, less intensive search efforts by females, less knowledge on the mortgage market by females, and weaker bargaining power by female borrowers. Due to the limitations of the data, we were not able to test and identify the source. Though we could not decisively conclude that the gender disparity found in mortgage lending is due to deliberate lender discrimination, we could not exclude this source of gender inequality. A clear conclusion of this study is women pay higher mortgage contract rates holding loan, property, neighborhoods, and other borrower financial characteristics constant.

References

- Agier, Isabelle, and Ariane Szafarz. "Microfinance and gender: is there a glass ceiling on loan size?" *World Development* 42 (2013): 165-181.
- Ahmed, Ali M., and Mats Hammarstedt. "Discrimination in the rental housing market: A field experiment on the Internet." *Journal of Urban Economics* 64, no. 2 (2008): 362-372.
- Ahmed, Ali M., Lina Andersson, and Mats Hammarstedt. "Are lesbians discriminated against in the rental housing market? Evidence from a correspondence testing experiment." *Journal of Housing Economics* 17, no. 3 (2008): 234-238.
- Alesina, Alberto F., Francesca Lotti, and Paolo Emilio Mistrulli. "Do women pay more for credit? Evidence from Italy." *Journal of the European Economic Association* 11, no. s1 (2013): 45-66.
- Allison, Paul D. *Survival analysis using SAS: a practical guide*. SAS Institute, 2010.
- Altonji, Joseph G., and Rebecca M. Blank. "Race and gender in the labor market." *Handbook of labor economics* 3 (1999): 3143-3259.
- Ayres, Ian. "Fair driving: Gender and race discrimination in retail car negotiations." *Harvard Law Review* (1991): 817-872.
- Ayres, Ian, and Peter Siegelman. "Race and gender discrimination in bargaining for a new car." *The American Economic Review* (1995): 304-321.
- Barber, Brad M., and Terrance Odean. "Boys will be boys: Gender, overconfidence, and common stock investment." *Quarterly journal of Economics* (2001): 261-292.
- Becker, Gary S. *The economics of discrimination*. University of Chicago press, 2010.
- Bellucci, Andrea, Alexander Borisov, and Alberto Zazzaro. "Does gender matter in bank-firm relationships? Evidence from small business lending." *Journal of Banking & Finance* 34, no. 12 (2010): 2968-2984.
- Black, Dan A., Amelia M. Haviland, Seth G. Sanders, and Lowell J. Taylor. "Gender wage disparities among the highly educated." *Journal of human resources* 43, no. 3 (2008): 630-659.
- Black, Harold, Robert L. Schweitzer, and Lewis Mandell. "Discrimination in mortgage lending." *The American Economic Review* 68, no. 2 (1978): 186-191.
- Blanchard, Lloyd, Bo Zhao, and John Yinger. "Do lenders discriminate against minority and woman entrepreneurs?" *Journal of Urban Economics* 63, no. 2 (2008): 467-497.
- Blanchflower, David G., Phillip B. Levine, and David J. Zimmerman. "Discrimination in the small-business credit market." *Review of Economics and Statistics* 85, no. 4 (2003): 930-943.

- Cavalluzzo, Ken S., and Linda C. Cavalluzzo. "Market structure and discrimination: The case of small businesses." *Journal of Money, Credit and Banking* (1998): 771-792.
- Cavalluzzo, Ken S., John D. Wolken, and Linda C. Cavalluzzo. "Competition, Small Business Financing, and Discrimination: Evidence from a New Survey." *Journal of Business* (2002): 641-679.
- Cheng, Ping, Zhenguo Lin, and Yingchun Liu. "Do women pay more for mortgages?" *The Journal of Real Estate Finance and Economics* 43, no. 4 (2011): 423-440.
- Cheng, Ping, Zhenguo Lin, and Yingchun Liu. "Racial discrepancy in mortgage interest rates." *The Journal of Real Estate Finance and Economics* 51, no. 1 (2015): 101-120.
- Coleman, Susan. "Access to capital and terms of credit: A comparison of men-and women-owned small businesses." *Journal of Small Business Management* 38, no. 3 (2000): 37.
- Cox, John C., Jonathan E. Ingersoll Jr, and Stephen A. Ross. "A theory of the term structure of interest rates." *Econometrica: Journal of the Econometric Society* (1985): 385-407.
- Deng, Yongheng, John M. Quigley, and Robert Order. "Mortgage terminations, heterogeneity and the exercise of mortgage options." *Econometrica* 68, no. 2 (2000): 275-307.
- Duan, Jin-Chuan, and Jean-Guy Simonato. "Estimating and testing exponential-affine term structure models by Kalman filter." *Review of Quantitative Finance and Accounting* 13, no. 2 (1999): 111-135.
- Duca, John V., and Stuart S. Rosenthal. "Do mortgage rates vary based on household default characteristics? Evidence on rate sorting and credit rationing." *The Journal of Real Estate Finance and Economics* 8, no. 2 (1994): 99-113.
- Eckel, Catherine, Angela De Oliveira, and Philip J. Grossman. "Gender and negotiation in the small: are women (perceived to be) more cooperative than men?" *Negotiation Journal* 24, no. 4 (2008): 429-445.
- Fishbein, Allen J., and Patrick Woodall. "Women are prime targets for subprime lending." *Consumer Federation of America* (2006).
- Frölich, Markus. "Propensity score matching without conditional independence assumption—with an application to the gender wage gap in the United Kingdom." *The Econometrics Journal* 10, no. 2 (2007): 359-407.
- Ghent, Andra C., Rubén Hernández-Murillo, and Michael T. Owyang. "Differences in subprime loan pricing across races and neighborhoods." *Regional Science and Urban Economics* 48 (2014): 199-215.
- Goldberg, Pinelopi Koujianou. "Dealer price discrimination in new car purchases: Evidence from the consumer expenditure survey." *Journal of Political Economy* (1996): 622-654.

- Kau, James B., Donald C. Keenan, Walter J. Muller III, and James F. Epperson. "The valuation at origination of fixed-rate mortgages with default and prepayment." *The Journal of Real Estate Finance and Economics* 11, no. 1 (1995): 5-36.
- Kau, James B., Donald C. Keenan, and Henry J. Munneke. "Racial discrimination and mortgage lending." *The Journal of Real Estate Finance and Economics* 45, no. 2 (2012): 289-304.
- Ladd, Helen F. "Equal credit opportunity: women and mortgage credit." *The American Economic Review* 72, no. 2 (1982): 166-170.
- Marianne, Bertrand. "New perspectives on gender." *Handbook of labor economics* 4 (2011): 1543-1590.
- Mayer, Chris, Tomasz Piskorski, and Alexei Tchistyi. "The inefficiency of refinancing: Why prepayment penalties are good for risky borrowers." *Journal of Financial Economics* 107, no. 3 (2013): 694-714.
- Morton, Fiona Scott, Florian Zettelmeyer, and Jorge Silva-Risso. "Consumer information and discrimination: Does the internet affect the pricing of new cars to women and minorities?" *Quantitative Marketing and Economics* 1, no. 1 (2003): 65-92.
- Munnell, Alicia H., Geoffrey MB Tootell, Lynn E. Browne, and James McEneaney. "Mortgage lending in Boston: Interpreting HMDA data." *The American Economic Review* (1996): 25-53.
- Niederle, Muriel, and Lise Vesterlund. "Do women shy away from competition? Do men compete too much?" *The Quarterly Journal of Economics* (2007): 1067-1101.
- Ñopo, Hugo. "Matching as a tool to decompose wage gaps." *The review of economics and statistics* 90, no. 2 (2008): 290-299.
- Turner, Margery Austin and Skidmore, Felicity (Editors). "Mortgage Lending Discrimination: A Review of Existing Evidence." The Urban Institute, 1999.
- Vandegrift, Donald, and Abdullah Yavas. "Men, women, and competition: An experimental test of behavior." *Journal of Economic Behavior & Organization* 72, no. 1 (2009): 554-570.
- Woodward, Susan E. "A study of closing costs for FHA mortgages." *US Department of Housing and Urban Development, Office of Policy Development and Research* (2008).
- Zhang, Yan. "Fair lending analysis of mortgage pricing: does underwriting matter?" *The Journal of Real Estate Finance and Economics* 46, no. 1 (2013): 131-151.

Table 2.1 Descriptive Statistics of Mortgage Loans at Loan Origination

Variable Name	Mean	Std. Dev.	Min.	Max.
Default (0,1)	0.0943	0.2923	0.0000	1.0000
Prepay (0,1)	0.7996	0.4004	0.0000	1.0000
Loan Characteristics				
Contract rate at origination (C_0)	8.0421	1.1704	5.2500	12.5000
10-year treasury yield at origination (y_0)	4.9580	0.8083	3.3300	6.8900
Contract rate spread at origination ($C_0 - y_0$)	3.0842	0.9829	0.8000	6.4750
Original LTV	84.8930	12.0897	35.0000	107.0000
FICO at origination	696.6768	56.6503	504.0000	822.0000
Original loan amount (in \$10,000)	16.9809	11.5623	2.0000	80.0000
Full income documentation (0,1)	0.4796	0.4997	0.0000	1.0000
Without prepayment penalty (0,1)	0.6428	0.4793	0.0000	1.0000
Prepayment penalty for 1 year (0,1)	0.0118	0.1079	0.0000	1.0000
Prepayment penalty for 2 years (0,1)	0.0136	0.1158	0.0000	1.0000
Prepayment penalty for 3 years (0,1)	0.2180	0.4130	0.0000	1.0000
Prepayment penalty for 5 years (0,1)	0.1138	0.3176	0.0000	1.0000
Property Characteristics				
Property owner occupied (0,1)	0.8518	0.3554	0.0000	1.0000
Property condo (0,1)	0.3513	0.4775	0.0000	1.0000
Neighborhood-Level Characteristics at Origination ^a				
Recent housing price appreciation rate ^b	0.1806	0.1969	-0.4952	2.3422
Heterogeneity in housing price (in \$10,000) ^c	10.8652	7.9879	1.1012	43.4989
Housing occupancy rate	0.9149	0.0808	0.5660	0.9970
Average household income ^d	1.1753	0.5985	0.3501	4.3600
Poverty rate	0.1380	0.0778	0.0219	0.5988
Proportion of African Americans	0.1418	0.2155	0.0006	0.9560
Sample Size	2,206			

^a The neighborhood of the property that secures a loan is defined as the property's location based on the 1990 census tract boundaries.

^b Recent housing price appreciation rate at origination is defined as the ratio of the median housing sale price in a census tract in a three-year period prior to the month of loan origination to the median housing sale price in the same census tract in another three-year period prior to the three-year pre-origination period, then minus 1.

^c Heterogeneity in housing price at origination is defined as the standard deviation of the housing sale price in a census tract over a three-year period prior to the month of loan origination.

^d Average household income at origination is defined as ratio of the average household income in a census tract at origination to the average household income in Miami-Dade County, FL at origination.

Table 2.2 Descriptive Statistics of Mortgage Loans at Loan Origination by Gender

Variable Name	Gender Group		
	A male sole borrower	A female sole borrower	Joint male and female co-borrowers
	Mean	Mean	Mean
Default (0,1)	0.1111	0.1464	0.0679
Prepay (0,1)	0.7820	0.7448	0.8273
Loan Characteristics			
Contract rate at origination (C_0)	8.0579	8.1616	7.9905
10-year treasury yield at origination (y_0)	4.9199	4.8804	5.0022
Contract rate spread at origination ($C_0 - y_0$)	3.1380	3.2812	2.9883
Original LTV	86.3082	86.0418	83.9145
FICO at origination	698.0440	691.9331	697.9680
Original loan amount (in \$10,000)	16.4199	14.4537	18.1604
Full income documentation (0,1)	0.4507	0.4163	0.5148
Without prepayment penalty (0,1)	0.6415	0.5858	0.6651
Prepayment penalty for 1 year (0,1)	0.0168	0.0146	0.0088
Prepayment penalty for 2 years (0,1)	0.0252	0.0126	0.0096
Prepayment penalty for 3 years (0,1)	0.1992	0.2594	0.2094
Prepayment penalty for 5 years (0,1)	0.1174	0.1276	0.1071
Property Characteristics			
Property owner occupied (0,1)	0.8008	0.8536	0.8705
Property condo (0,1)	0.4151	0.4498	0.2894
Neighborhood-Level Characteristics at Origination ^a			
Recent housing price appreciation rate ^b	0.1843	0.1962	0.1732
Heterogeneity in housing price (in \$10,000) ^c	11.1721	9.6551	11.2106
Housing occupancy rate	0.8962	0.9142	0.9224
Average household income ^d	1.1497	1.0368	1.2380
Poverty rate	0.1554	0.1473	0.1279
Proportion of African Americans	0.1552	0.1627	0.1287
Number of Loans	477	478	1,251

^a The neighborhood of the property that secures a loan is defined as the property's location based on the 1990 census tract boundaries.

^b Recent housing price appreciation rate at origination is defined as the ratio of the median housing sale price in a census tract in a three-year period prior to the month of loan origination to the median housing sale price in the same census tract in another three-year period prior to the three-year pre-origination period, then minus 1.

^c Heterogeneity in housing price at origination is defined as the standard deviation of the housing sale price in a census tract over a three-year period prior to the month of loan origination.

^d Average household income at origination is defined as ratio of the average household income in a census tract at origination to the average household income in Miami-Dade County, FL at origination.

Table 2.3 Loan Hazard Model Estimates

Variables	Model (1) Default		Model (2) Prepay	
	Coef.	P-value	Coef.	P-value
Intercept	-1.1213	0.7760	-11.7734	<0.0001
Contract rate spread at origination ($C_0 - y_0$)	0.2696	0.2873	0.5478	<0.0001
Market interest rate change at time t ($y_0 - y_{t-2}$) ^a	0.3104	0.4255	0.5592	<0.0001
Relative house price at time t (RHP_t) ^a	-2.1651	<0.0001	1.2509	<0.0001
$(y_0 - y_{t-2}) \times (RHP_t)$ ^a	0.2249	0.5112	-0.0350	0.6916
Original LTV categories (Base group: Original LTV ≤ 75)				
75 < Original LTV ≤ 80	0.5595	0.0530		
80 < Original LTV ≤ 90	0.6438	0.0626		
90 < Original LTV ≤ 100	0.8593	0.0261		
100 < Original LTV	1.7259	0.0014		
FICO at origination continuous linear splines				
Minimum (FICO, 700)	-0.0094	0.0289	0.0056	<0.0001
Maximum (FICO, 700) - 700	-9.26E-05	0.7163	-4.61E-05	0.5966
Original loan amount (in \$10,000)	0.1887	<0.0001	0.0228	0.0074
Square term of original loan amount (in \$10,000)	-0.0018	<0.0001	-0.0003	0.0472
Full income documentation (0,1)	-0.7481	<0.0001		
Within prepayment penalty period at time t (0,1) ^a			-0.6270	<0.0001
Property owner occupied (0,1)	-0.7915	0.0005	0.0170	0.8265
Property condo (0,1)	0.5760	0.0054	-0.2813	<0.0001
Neighborhood-level heterogeneity in housing price at time t (in \$10,000) ^a	-0.0258	0.1630		
Neighborhood-level housing occupancy rate at time t ^a	0.2860	0.7993	-0.0917	0.7906
Neighborhood-level average household income at time t ^a	-0.9383	0.0018	0.1534	0.0095
Neighborhood-level poverty rate at time t ^a	-0.6473	0.6477	-0.2058	0.6671
Neighborhood-level proportion of African Americans at time t ^a	1.2208	0.0007	-0.0241	0.8684
Borrower gender categories (Base group: Joint male and female co-borrowers)				
Male sole borrower (0,1)	0.5287	0.0594	0.0656	0.5271
Female sole borrower (0,1)	0.6019	0.0361	-0.0614	0.5729
Married (0,1)	-0.0807	0.7611	-0.0133	0.8925
Residual ^b	0.3538	0.1935	-0.0918	0.2849
Baseline hazard ^a				
SDA	1.5269	<0.0001		
Mortgage year fixed effects ^c			YES	
Origination season fixed effects ^c	YES		YES	
Likelihood Ratio	391.7133	<0.0001	1024.1004	<0.0001

^a Demote time-varying variables.^b The residual comes from the contract-rate reduced-form estimation.^c Loan origination season fixed effects estimates and mortgage year fixed effects estimates are omitted here, but are available upon request.

Table 2.4 Loan Contract Rate Estimates -2SLS

Variables	Model (1) ^a		Model (2) ^b	
	Coef.	P-value ^c	Coef.	P-value ^c
Intercept	11.6145	<0.0001	11.3919	<0.0001
Predicted default probability	1.2489	<0.0001	1.3787	<0.0001
Predicted pecuniary prepayment probability	2.1714	<0.0001	2.6539	<0.0001
Predicted non-pecuniary prepayment probability	1.3600	<0.0001	1.6981	<0.0001
10-year treasury yield at origination (y_0)	0.6634	<0.0001	0.6795	<0.0001
Original LTV categories (Base group: Original LTV<=75)				
75<Original LTV<=80	0.0585	0.0839	0.0547	0.1062
80<Original LTV<=90	0.2551	<0.0001	0.2526	<0.0001
90<Original LTV<=100	0.3998	<0.0001	0.3951	<0.0001
100<Original LTV	0.6936	<0.0001	0.6751	<0.0001
FICO at origination continuous linear splines				
Minimum (FICO, 700)	-0.0116	<0.0001	-0.0115	<0.0001
Maximum (FICO, 700) -700	4.91E-05	0.2427	5.18E-05	0.2212
Original loan amount (in \$10,000)	-0.0284	<0.0001	-0.0287	<0.0001
Square term of original loan amount (in \$10,000)	0.0003	<0.0001	0.0003	<0.0001
Full income documentation (0,1)	-0.0801	0.0084	-0.0743	0.0138
Prepayment penalty categories (Base group: No prepayment penalty)				
Prepayment penalty for 1 year	0.7574	<0.0001	0.7320	<0.0001
Prepayment penalty for 2 years	0.5622	<0.0001	0.5364	<0.0001
Prepayment penalty for 3 years	0.6225	<0.0001	0.6101	<0.0001
Prepayment penalty for 5 years	0.4685	<0.0001	0.4573	<0.0001
Property owner occupied (0,1)	-0.0993	0.0052	-0.0970	0.0066
Property condo (0,1)	0.0704	0.0675	0.0713	0.0646
Neighborhood-level recent housing price appreciation rate at origination	-0.2467	0.0075	-0.2375	0.0098
Neighborhood-level heterogeneity in housing price at origination (in \$10,000)	-0.0031	0.2295	-0.0032	0.2261
Neighborhood-level housing occupancy rate at origination	-0.2256	0.2432	-0.2101	0.2770
Neighborhood-level average household income at origination	-0.0785	0.0460	-0.0702	0.0771
Neighborhood-level poverty rate at origination	0.2682	0.2519	0.2684	0.2543
Neighborhood-level proportion of African Americans at origination	0.0388	0.6068	0.0337	0.6565
Borrower gender categories (Base group: Joint male and female co-borrowers)				
Male sole borrower (0,1)	0.0538	0.2951	0.0490	0.3414
Female sole borrower (0,1)	0.1315	0.0108	0.1280	0.0134
Married (0,1)	0.0688	0.1438	0.0672	0.1549
Time Trend	-0.0129	0.1403	-0.0084	0.3711
Loan origination season fixed effects ^d	YES		YES	
Adjusted R ²	0.7737		0.7721	

^a In model (1), pecuniary prepayment is defined as occurring when the future 10-year yield at time t drops below the yield at origination, while non-pecuniary prepayment occurs when the future 10-year yield at time t is above the yield at origination.

^b In model (2), pecuniary prepayment is defined as occurring when the future 10-year yield at time t drops by more than 100 basis points relative to the yield at origination, and non-pecuniary prepayment occurs when the 10-year yield at time t exceeds the yield at origination, or drops by less than 100 basis points.

^c P-values were calculated using standard-error corrections because predicted termination probabilities were generated variables. The Standard way to correct the standard error demonstrated in Appendix 6A of Wooldridge (2010) was employed.

^d Loan origination season fixed effects estimates are omitted here, but are available upon request.

Table 2.5 Loan Contract Rate Estimates with Gender-Race Interaction -2SLS

Variable	Coef.	P-value ^a
Intercept	11.6166	<0.0001
Predicted default probability	1.2455	<0.0001
Predicted pecuniary prepayment probability ^b	2.1727	<0.0001
Predicted non-pecuniary prepayment probability ^b	1.3621	<0.0001
10-year treasury yield at origination (y_0)	0.6631	<0.0001
Original LTV categories (Base group: Original LTV ≤ 75)		
75 < Original LTV ≤ 80	0.0599	0.0755
80 < Original LTV ≤ 90	0.2561	<0.0001
90 < Original LTV ≤ 100	0.4014	<0.0001
100 < Original LTV	0.6977	<0.0001
FICO at origination continuous linear splines		
Minimum (FICO, 700)	-0.0116	<0.0001
Maximum (FICO, 700) - 700	4.93E-05	0.2399
Original loan amount (in \$10,000)	-0.0284	<0.0001
Square term of original loan amount (in \$10,000)	0.0003	<0.0001
Full income documentation (0,1)	-0.0818	0.0072
Prepayment penalty categories (Base group: No prepayment penalty)		
Prepayment penalty for 1 year	0.7525	<0.0001
Prepayment penalty for 2 years	0.5631	<0.0001
Prepayment penalty for 3 years	0.6217	<0.0001
Prepayment penalty for 5 years	0.4689	<0.0001
Property owner occupied (0,1)	-0.1004	0.0046
Property condo (0,1)	0.0707	0.0656
Neighborhood-level recent housing price appreciation rate at origination	-0.2479	0.0076
Neighborhood-level heterogeneity in housing price at origination (in \$10,000)	-0.0032	0.2205
Neighborhood-level housing occupancy rate at origination	-0.2199	0.2560
Neighborhood-level average household income at origination	-0.0775	0.0490
Neighborhood-level poverty rate at origination	0.2804	0.2366
Neighborhood-level proportion of African Americans at origination	0.0492	0.5548
Borrower gender categories (Base group: Joint male and female co-borrowers)		
Male sole borrower (0,1)	0.0594	0.2496
Male sole borrower (0,1) × African American Neighborhood (0,1) ^c	-0.0813	0.4454
Female sole borrower (0,1)	0.1247	0.0169
Female sole borrower (0,1) × African American Neighborhood (0,1) ^c	0.0349	0.7218
Married (0,1)	0.0658	0.1613
Time Trend	-0.0131	0.1341
Loan origination season fixed effects ^d	YES	
Adjusted R ²	0.7737	

^a P-values were calculated using standard-error corrections because predicted termination probabilities were generated variables. Standard way to correct the standard error demonstrated in Appendix 6A of Wooldridge (2010) was employed.

^b It is defined that pecuniary prepayment occurs when the future 10-year yield at time t drops below the yield at origination, while non-pecuniary prepayment occurs when the future 10-year yield at time t is above the yield at origination.

^c A neighborhood is defined as a predominantly African American one if more than 50% of the residents in that neighborhood are African Americans. The results are consistent with alternative definitions of a predominantly African American neighborhood. With alternative definitions, a neighborhood is defined an African American one if more than 60% (or more than 70%) of the residents in that neighborhood are African Americans. The results with alternative definitions are omitted here, but are available upon request.

^d Results on loan origination season fixed effects are omitted here, but are available upon request.

Table 2.6 Nonparametric Matching Results on Nearest 1-to-1 Matching with Replacements and with Calipers ^a

Variable	Matching Sample 1 ^c						Matching Sample 2 ^c					
	Calipers: $\hat{P}_k$: $\pm 5\%$; <i>Orig. Month</i> : ± 18 mths; <i>Dist</i> : N/A						Calipers: $\hat{P}_k$: $\pm 5\%$; <i>Orig. Month</i> : ± 18 mths; <i>Dist</i> : 7.5 mi					
	Mean (Female)	Mean (Male)	Diff. of the mean	Std. diff. of the mean	Ratio of the var.	Mean diff. <i>t</i> test	Mean (Female)	Mean (Male)	Diff. of the mean	Std. diff. of the mean	Ratio of the var.	Mean diff. <i>t</i> test
Contract rate at origination	8.1739	7.8736	0.3003	0.2692	1.2557	4.0200	8.1992	7.9590	0.2401	0.2107	1.2306	2.7900
Matching Balance Diagnostic Variables ^b												
10-year treasury yield at origination	4.8920	4.8853	0.0067	0.0086	1.0405	0.1300	4.9231	4.9114	0.0117	0.0145	1.0580	0.1900
Predicted default probability	0.1093	0.1082	0.0011	0.0088	0.9873	0.1300	0.0939	0.0915	0.0024	0.0214	1.0131	0.2800
Predicted pecuniary prepayment probability	0.3242	0.3252	-0.0010	-0.0056	0.9970	-0.0800	0.3363	0.3400	-0.0036	-0.0200	1.0152	-0.2600
Predicted non-pecuniary prepayment probability	0.3106	0.3135	-0.0029	-0.0216	1.0077	-0.3200	0.3154	0.3195	-0.0041	-0.0307	1.0051	-0.4100
Number of Match Pairs	447						350					

**Table 2.6 Nonparametric Matching Results on Nearest 1-to-1 Matching with Replacements and with Calipers
(Continued)**

Variable	Matching Sample 3 ^c						Matching Sample 4 ^c					
	Calipers: $\hat{P}_k : \pm 2.5\%$; <i>Orig. Month</i> : ± 18 mths; <i>Dist</i> : N/A						Calipers: $\hat{P}_k : \pm 2.5\%$; <i>Orig. Month</i> : ± 18 mths; <i>Dist</i> : 7.5 mi					
	Mean (Female)	Mean (Male)	Diff. of the mean	Std. diff. of the mean	Ratio of the var.	Mean diff. <i>t</i> test	Mean (Female)	Mean (Male)	Diff. of the mean	Std. diff. of the mean	Ratio of the var.	Mean diff. <i>t</i> test
Contract rate at origination	8.2699	7.9851	0.2848	0.2553	1.1380	3.1700	8.4662	8.1166	0.3495	0.2995	1.2825	2.6700
Matching Balance Diagnostic Variables ^b												
10-year treasury yield at origination	4.9871	4.9733	0.0139	0.0170	1.0290	0.2100	5.1547	5.1318	0.0228	0.0266	1.0143	0.2400
Predicted default probability	0.0862	0.0855	0.0008	0.0071	0.9880	0.0900	0.0737	0.0723	0.0014	0.0158	1.0054	0.1400
Predicted pecuniary prepayment probability	0.3568	0.3577	-0.0009	-0.0051	0.9990	-0.0600	0.3955	0.3952	0.0003	0.0017	1.0102	0.0200
Predicted non-pecuniary prepayment probability	0.3138	0.3166	-0.0028	-0.0198	0.9848	-0.2500	0.2970	0.2981	-0.0011	-0.0083	0.9895	-0.0700
Number of Match Pairs	305						159					

^a Nearest 1-to-1 matching with replacements and with calipers is used to match loans of a female sole borrower to loans of a male sole borrower in the sample. For each loan of a female sole borrower, all the loans of a male sole borrower that satisfy the calipers are selected as potential “matches”. A caliper is used to impose a tolerance level for the maximum distance on a matching criterion between a loan of a female borrower and a loan of a male borrower. The matching criteria include the predicted probability of each event (default, pecuniary prepayment, and non-pecuniary prepayment), loan origination month, and the geographic location of the property securing the loan. For each loan of a female borrower, within the resulting set of potential “matches”, the loan with the smallest total difference in loan termination probability is selected as the sole and best match. The total difference in loan termination probability is defined as the quadratic mean of the differences in the predicted default probability ($k=1$), the predicted pecuniary prepayment probability ($k=2$), and the predicted non-pecuniary prepayment probability ($k=3$) between a female borrower and a male borrower. Specifically, it is defined as follows: $|\hat{P}^{female} - \hat{P}^{male}| = \sqrt{\sum_{k=1}^3 (\hat{P}_k^{female} - \hat{P}_k^{male})^2 / 3}$, where $\hat{P}_k^{female}$ is the predicted probability of event k for a female borrower, and $\hat{P}_k^{male}$ is the predicted probability of event k for a male borrower.

^b In order to check whether the female group and the male group in the resulting matching sample have the same distributions of loan termination risk and baseline mortgage interest rate at origination, four variables are used to diagnose whether the matching sample is well balanced. They are the 10-year treasury constant maturity yield at origination and the predicted probability of each event k . The standardized difference of the mean and the ratio of the variance of each balance diagnostic variable between the female group and the male group are used to measure whether each matching sample is well balanced following the guidelines in Rubin (2001).

^c The four matching samples differ by the specifications of the calipers. Specifications of the calipers vary by the maximum difference acceptable for matching on the predicted probability of each event k ($k=1, 2, 3$) between a loan of a female sole borrower and a loan of a male sole borrower, as well as by whether a caliper on the distance of the geographic location of the property securing a loan is applied. For matching sample 1 and 2, it is required the difference in the predicted probability of each event k between a loan of a female sole borrower and a loan of a male sole borrower be less than or equal to 5%, while for matching sample 3 and 4, the difference in the predicted probability of each event k should be less than or equal to 2.5%. Meanwhile, for matching sample 1 and 3, the location distance caliper is not applied, while for matching sample 2 and 4, the location distance caliper is applied, with a maximum distance of 7.5 miles. Location distance calipers of 10 miles and 12.5 miles are also applied respectively. The results are consistent and are omitted here, but are available upon request. Additionally, a caliper on loan origination month distance is applied for all of the matching samples, with a maximum distance of 18 months.

Appendices

Table A.1 Definitions of Variables

Variables	Variable Definition
Loan Characteristics	
Contract rate at origination (C_0)	Contract interest rate at origination, in percentage
10-year treasury yield at origination (y_0)	10-year treasury constant maturity yield at origination, in percentage
Contract rate spread at origination ($C_0 - y_0$)	Contract interest rate at origination minus 10-year treasury constant maturity yield at origination, in percentage
Original LTV	The LTV ratio at loan origination, in percentage
FICO at origination	Credit score at loan origination
Original loan amount (in \$10,000)	Loan size at loan origination (in \$10,000)
Full income documentation (0,1)	=1 for a loan with full documentation of a borrower's income
Prepayment penalty for m year(s) ($m=1,2,3,5$)	=1 for a loan encumbered by a prepayment penalty for the first m year(s)
Within prepayment penalty period (0,1) ^a	=1 for a loan encumbered by a prepayment penalty in a given mortgage month t
Property Characteristics	
Property owner occupied (0,1)	=1 for a loan secured by a property occupied by the owner
Property condo (0,1)	=1 for a loan secured by a condo
Relative house price at time t (RHP) _{t} ^a	The ratio of the median house sale price index at time t to the median house sale price index at loan origination ^b
Borrower Characteristics	
Male sole borrower (0,1)	=1 for a loan signed by one male borrower
Female sole borrower (0,1)	=1 for a loan signed by one female borrower
Joint male and female co-borrowers (0,1)	=1 for a loan co-signed by two borrowers, one is a male and the other is a female
Married (0,1)	=1 for a loan signed by (a) borrower(s) who are married at loan origination

Table A.1 Definitions of Variables (Continued)

Variables	Variable Definition
Neighborhood-Level Characteristics	
Recent housing price appreciation rate at origination	The ratio of the median housing sale price in a census tract in a three-year period prior to the month of loan origination to the median housing sale price in the same census tract in another three-year period prior to the three-year pre-origination period, then minus 1
Heterogeneity in housing price ^c	The standard deviation of the housing sale price in a census tract over a three-year period prior to a given month (in \$10,000)
Housing occupancy rate ^c	Housing occupancy rate in a census tract
Average household income ^c	Ratio of the average household income in a census tract to the average income in Miami-Dade County, FL
Poverty rate ^c	Poverty rate in a census tract
Proportion of African Americans ^c	Proportion of households who are African Americans in a census tract
Interest Rate Environment	
Market interest rate change at time t ($y_0 - y_{t-2}$) ^a	10-year treasury yield at loan origination minus 10-year treasury yield at time t , lagged by 2 months, in percentage
Others	
Time Trend	A variable that equals to 1 for a loan originated in year 1997, 2 for a loan originated in year 1998, and so forth

^a Denotes time-varying variables in default/prepayment hazard model.

^b This median house price index is at the census tract level, and is generated from the property transaction data base in Miami-Dade County, FL. For each month, this index is calculated based on the inflation-adjusted median housing sale price over a three-year window around that month, eighteen months before and eighteen months after that given month. All prices are defined in 2009 dollars.

^c Denotes variables that are time-varying variables in default/prepayment hazard model, but represent values at loan origination in the contract rate determination model.

Table A.2 Loan Contract Rate Reduced-Form Estimates

Variable	Coef.	P-value
Intercept	12.5492	<0.0001
10-year treasury yield at origination (y ₀)	0.8160	<0.0001
Original LTV categories (Base group: Original LTV≤75)		
75<Original LTV≤80	0.0989	0.0260
80<Original LTV≤90	0.3291	<0.0001
90<Original LTV≤100	0.4867	<0.0001
100<Original LTV	0.8433	<0.0001
FICO at origination continuous linear splines		
Minimum (FICO, 700)	-0.0119	<0.0001
Maximum (FICO, 700) -700	3.35E-05	0.5109
Original loan amount (in \$10,000)	-0.0218	<0.0001
Square term of original loan amount (in \$10,000)	0.0003	0.0004
Full income documentation (0,1)	-0.1122	0.0013
Prepayment penalty categories (Base group: No prepayment penalty)		
Prepayment penalty for 1 year	0.8807	<0.0001
Prepayment penalty for 2 years	0.6131	<0.0001
Prepayment penalty for 3 years	0.6289	<0.0001
Prepayment penalty for 5 years	0.3400	<0.0001
Property owner occupied (0,1)	-0.1064	0.0231
Property condo (0,1)	-0.0196	0.6434
Neighborhood-level recent housing price appreciation rate at origination	-0.3787	<0.0001
Neighborhood-level heterogeneity in housing price at origination (in \$10,000)	-0.0053	0.1084
Neighborhood-level housing occupancy rate at origination	-0.3397	0.1664
Neighborhood-level average household income at origination	-0.0402	0.3869
Neighborhood-level poverty rate at origination	0.4426	0.1041
Neighborhood-level proportion of African Americans at origination	0.0763	0.3675
Borrower gender categories (Base group: Joint male and female co-borrowers)		
Male sole borrower (0,1)	0.1234	0.0409
Female sole borrower (0,1)	0.1609	0.0091
Married (0,1)	0.0782	0.1696
Time Trend	-0.0422	<0.0001
Loan origination season fixed effects ^a	YES	
Adjusted R ²	0.6836	

^a Results on loan origination season fixed effects are omitted here, but are available upon request.

CHAPTER 3

ESSAY TWO: DON'T FORGET ABOUT DEFAULT AND PREPAYMENT – AN ANALYSIS OF MORTGAGE LENDING³⁵

³⁵ Fang, Lu, James B. Kau, Henry J. Munneke. To be submitted to *The Journal of Real Estate Finance and Economics*.

Abstract

This paper investigates whether a borrower's race and the racial composition of a borrower's neighborhood impact loan contract rate beyond the extent to which those traits affect loan default probability and prepayment probability. In order to obtain a borrower's race information, this paper matches loans drawn from a pool of subprime mortgage loans to HMDA data. A competing-risks loan hazard model is estimated to examine the degree to which a borrower's race and the racial composition of a borrower's neighborhood are associated with loan default probability and prepayment probability based on the data on loan performance. The predicted loan default probability and prepayment probability are incorporated in a loan contract rate determination model to investigate whether the individual- and neighborhood-level race variables have an additional impact on the contract rate. A loan-level matching approach, based on the loan-level predicted default probability and prepayment probability, is utilized to test if there is a systematic difference in contract rate between two racial groups after any systematic differences in loan risk are eliminated. Both the regression results and the matching results reveal that African American borrowers and Hispanic borrowers tend to pay higher contract rates than their comparable White counterparts in the subprime mortgage market over the study period. The racial composition of a borrower's neighborhood is not shown to affect loan pricing.

3.1. Introduction

The issue of racial discrimination in mortgage lending has drawn attention since 1974 when the Equal Credit Opportunity Act (ECOA) was enacted (Turner & Skidmore, 1996; Ladd, 1998; LaCour-Little, 1999; Ross & Yinger, 2002).³⁶ Overall, prior studies have not arrived at an agreement on whether racial discrimination exists in mortgage lending. Therefore, this issue remains an open question. Racial discrimination in mortgage markets may manifest itself in several forms. Lenders may practice racial discrimination by denying loan applications more frequently by minority applicants, or by applicants living in a predominantly minority neighborhood. Lenders may also discriminate against minority borrowers by steering them into the subprime mortgage market, and finally by varying loan contract rates simply based on a borrower's race or the racial composition of a borrower's neighborhood. This study focuses on the second form of racial discrimination described above – lenders' loan pricing behavior, given the shift of the whole mortgage lending industry away from a system in which lenders would not allocate credit to risky borrowers to the one where risk-based pricing becomes more prevalent since the late 1990s (Turner & Skidmore, 1999; Ghent et al., 2014). The rise of the subprime mortgage market may have provided credit to risky borrowers, but it also might have allowed for charging at different contract rates.

This study examines whether mortgage lenders charge different loan contract rates based on a borrower's race (ethnicity) or the racial (ethnic) composition of a borrower's neighborhood beyond the extent to which those factors impact loan termination patterns. In this study, the loan data set consists of 30-year first-lien fixed-rate subprime mortgage loans for home purchase originated in Miami-Dade County, FL from 1997 to 2006. These data are matched with Home

³⁶ Turner & Skidmore (1996), Ladd (1998), LaCour-Little (1999), and Ross & Yinger (2002) provided thorough and detailed reviews of previous studies on racial discrimination in the mortgage market.

Mortgage Disclosure Act (HMDA) data to obtain individual- and neighborhood-level racial (ethnic) information for each loan.

The existence of loan contract rate disparity across various racial (ethnic) groups is evaluated from a lender's perspective. For each loan in the sample, the probability of a borrower defaulting upon or prepaying a loan is predicted, using a competing-risks loan hazard model on the basis of the loan performance data. In the loan hazard model, the effects of a borrower's race (ethnicity) and the racial (ethnic) composition of a borrower's neighborhood on loan termination probabilities are assessed. The loan-level predicted default probability and prepayment probability are incorporated into a loan contract rate determination model to investigate whether racial disparity in loan contract rate is attributed to race or its impact on tendencies to default or prepay a loan. In addition to the regression approach, a matching technique is also applied to pair each loan in a particular racial (ethnic) group to a comparable loan in another group based on various risk factors including the predicted default probability and prepayment probability. This matching technique is employed to evaluate if there is a systematic disparity in loan contract rate between two racial (ethnic) groups when any systematic differences in loan termination risk are mitigated by matching.

The regression results on loan hazard model provide empirical evidence that a borrower's race (ethnicity) is associated with loan prepayment probability, while racial (ethnic) composition does not have any influence on either default probability or prepayment probability. The results demonstrate that African American borrowers and Hispanic borrowers tend to be less likely to prepay than similar non-Hispanic White borrowers. With this correlation accounted for in the loan contract rate determination model, the regression results based on the whole pooled sample show that a borrower's race (ethnicity) does have an additional impact on a loan's contract rate, while

the racial (ethnic) composition does not. Specifically, African American borrowers appear to pay contract rates approximately 30 basis points higher than non-Hispanic White counterparts, while Hispanic borrowers tend to pay contract rates nearly 10 basis points higher than non-Hispanic White borrowers. The matching results based on a matched sample of African American borrowers and non-Hispanic White borrowers seem very close to the regression results based on the whole pooled sample in terms of the magnitude of the racial difference in contract rate. However, the results on matching between Hispanic borrowers and non-Hispanic White borrowers indicate a wider contract rate disparity than what the regression results on the pooled sample indicate. The rate disparity between Hispanic borrowers and non-Hispanic White borrowers indicated by the matching technique ranges from 23 basis points to 30 basis points.

Several possible explanations for the observed racial (ethnic) disparity in loan contract rate are proposed and tested. Those explanations include discrimination and regulation. Regulation may contribute to the observed rate disparity because in this study we have found that African American borrowers and Hispanic borrowers are less likely to exercise the prepayment option for their loans, and thus should be charged at a corresponding lower contract rate in a competitive world, *ceteris paribus*; however, lenders, with the fear of regulation, would deliberately ignore the known effects of race (ethnicity) on loan termination patterns and charge the same contract rates to them as to the non-Hispanic White borrowers, keeping other things equal. A race-blind model is used to examine the extent to which regulation contributes to the observed rate disparity. In this race-blind model, predicted prepayment probabilities are calculated in a race-blind world where lenders are assumed to disregard the impact of race (ethnicity) on prepayment risk, and to treat African American borrowers and Hispanic borrowers as non-Hispanic White borrowers. The results on this “race-blind” model reveal that the observed rate disparity between Hispanic

borrowers and non-Hispanic White borrowers is entirely explained by regulation, and roughly half of the observed rate disparity between African American borrowers and non-Hispanic White borrowers is attributable to regulation, with the remaining part due to discrimination.

The methodology applied in this study makes this study noticeably different from prior studies. In most of the prior racial studies in mortgage lending, the existence of racial disparity in contract rate is evaluated through a reduced-form contract rate determination model. In a reduced-form model, a loan's contract rate is regressed against a borrower's race (ethnicity) or the racial (ethnic) composition of a borrower's neighborhood as well as a set of covariates that are believed to be associated with the risk level of a loan. One potential problem involved in this reduced-form model is it ignores the possibility that a borrower's race (ethnicity) or the racial (ethnic) composition might be correlated with some loan risk factors unobserved to us. If that is the case, the estimate on the race (ethnicity) variable, either the individual-level one or the neighborhood-level one, would represent the influence of those unobserved factors on loan contract rate rather than the pure impact of the race variable itself. By contrast, the methodologies employed here including both the regression analysis and the matching technique explicitly and completely account for loan termination risk either by including the predicted default probability and prepayment probability in a contract rate determination model or by matching loans in two racial (ethnic) groups based on those predicted termination probabilities.

Another noticeable difference between this study and prior studies is the current study considers the two competing termination risks – default risk and prepayment risk, and simultaneously accounts for the association between race (ethnicity) and default risk as well as the correlation between race (ethnicity) and prepayment risk; whereas prior studies normally ignore the role of prepayment risk in loan pricing and the possible correlation between race (ethnicity)

and prepayment patterns.³⁷ Several potential problems may arise if prepayment risk is not taken into account. Based on the option-based pricing model of fixed-rate mortgage loans (Kau, et al., 1995), as a mortgage loan is a complex financial instrument involving various contractual provisions that interact in a complex way including default option and prepayment option, the price of a loan could not be determined adequately without consideration of any of them. Meanwhile, several prior studies have found empirical evidence that a borrower's race (ethnicity) is associated with loan prepayment patterns (Kelly, 1995; Clapp et al, 2001; Deng and Gabriel, 2006; Firestone, et al., 2007; An et al., 2010).³⁸ These findings further support our intention of controlling for the correlation between race (ethnicity) and prepayment risk when examining the relationship between race (ethnicity) and contract rate. Additionally, including prepayment risk within the competing-risks analysis framework enables us to assess the extent to which race (ethnicity) is associated with prepayment rate, and thus allows us to use the “race-blind” model to identify whether and the extent to which regulation contributes to the contract rate disparity across racial (ethnic) groups found in this study. By contrast, the traditional reduced-form contract rate model or any model that fails to take prepayment risk into account could not identify this possible rate disparity source – regulation, and falsely concludes that the rate disparity (if observed) is wholly attributable to discrimination.

The paper is organized as follows. The second section offers a literature review. The third section explains the empirical model in detail. The fourth section describes the data. The fifth

³⁷ Ghent et al. (2014) is one of the few racial studies in mortgage lending that considers the possible correlation between race (ethnicity) and loan default risk, however, it does not take prepayment risk into account.

³⁸ Kelly (1995) using loan-level data from the VA's mortgage program from 1971 to 1989 found that African American borrowers and Hispanic borrowers prepaid less frequently than Whites. Based on a pool of fixed-rate residential mortgage loans originated from 1993 to 1994, with their performance tracked through 1998, Clapp et al. (2001) concluded that a borrower's minority status is negatively associated with prepayment rate. The same conclusions could also be found in Deng and Gabriel (2006), Firestone, et al. (2007), and An et al. (2010).

section explains model specifications and matching Algorithm. The sixth section presents the results. The seventh section discusses the results and final section provides the conclusion.

3.2. Literature Review

As this study focuses on racial inequality in loan pricing instead of loan application, literature review here is limited to prior studies that evaluate the racial (ethnic) disparity in mortgage loan contract rates, not disparity in the likelihood of a loan being approved by the lender.

Lenders may discriminate against minority borrowers either simply based on a borrower's race (ethnicity) or based on the racial composition of a borrower's neighborhood. The second scenario is referred as redlining. Overall, previous studies have not reached a consensus on whether racial disparity exists in mortgage lending. The lack of consensus may be due to different data sources and/or different modelling techniques employed.

The Survey of Consumer Finance (SCF) data which has information on a borrower's race (ethnicity) has been widely used in prior studies. Duca and Rosenthal (1994), using the 1983 SCF data and focusing on conventional fixed-rate mortgage loans, did not find any evidence of racial discrimination in the conventional mortgage market. Getter (2006), uses the 1998 and 2001 SCF data which better differentiates households with severe delinquency problems from households with minor repayment problems than the 1983 data, arrived at the same conclusion: a borrower's race does not affect a loan's contract rate. However, Cheng et al. (2014) using the 2001, 2004, and 2007 SCF data consisting of various types of mortgage products, find that racial disparity in mortgage contract rate is both statistically and economically significant. The newer SCF data used by Cheng et al. (2014) covers detailed information on a borrower's race as well as the borrower's characteristics including income, wealth, debts, credit quality, age, and education.

Another data source employed in racial studies in mortgage lending is the American Housing Survey (AHS) data. Based on the 1989-2001 AHS data, Boehm et al. (2010) employed the Blinder's decomposition approach, commonly used in empirical studies on discrimination in the labor market, to provide empirical evidence of racial discrimination in the mortgage market. Specifically, their results demonstrate that on average, African American borrowers pay an Annual Percentage Rate (APR) approximately 20 basis points higher for home purchase mortgage loans, and around 94 basis points higher for refinancing loans than white counterparts. Meanwhile, Hispanic borrowers are shown to pay approximately 12 basis points higher than white counterparts for home purchase loans, but they do not pay significantly higher APRs for refinancing loans.

Although both the SCF data and the AHS data appear to provide rich data sources for analysis of racial disparity in mortgage lending, they have some drawbacks. The SCF data (a borrower-level data set) does not have much information on the attributes of the loan nor the collateral, although it includes detailed information on the borrower. The AHS data which is household-level data, fails to provide information on a household's financial condition (e.g., a borrower's credit score), but it includes detailed information on the household, the loan terms, and the collateral. The limitations of these two data sources have led to the use of proprietary data sets to investigate the issue of racial disparity in mortgage lending, with the hope these data will contain information to better account for risk.

Based on a proprietary data set consisting of conventional loans and FHA/VA loans originated by a national home mortgage lender from 1988 to 1989, Crawford and Rosenblatt (1999) focus on differences in yield premiums across racial groups. They conclude that conventional loan interest rates are race-neutral, and although African American borrowers are shown to pay a smaller premium on average, the difference is not economically significant. Black et al. (2003)

utilized another proprietary data set from a major mortgage lending institution to investigate whether racial differences in overage pricing exist in either home purchase mortgage market or refinancing market.³⁹ Their results mainly indicate that minority borrowers are significantly more likely to pay a positive overage than similarly situated white borrowers for home purchase mortgage loans, and when they do, they tend to pay a higher overage. However, they did not find any significant difference in overage pricing across racial groups in the refinancing mortgage market. Courchane (2007), using a sample of conventional home purchase loans and refinancing loans, failed to provide any evidence of racial discrimination in either of the two mortgage markets. By contrast, Zhang (2013), matched a proprietary loan data from a national bank to HMDA data to obtain information on a borrower's race. This study concludes that non-Hispanic Asians and non-Hispanic African Americans tend to receive a lower price for their first-lien conventional loans for home purchase.

Other studies have investigated the issue of racial discrimination in mortgage lending by emphasizing on the issues of redlining. Nothaft and Perry (2002), matched their loan origination data to census survey data to obtain the racial composition of a borrower's neighborhood. They find that borrowers in predominately Hispanic and Asian neighborhoods pay slightly higher interest rates, while borrowers in predominately African American neighborhoods occasionally pay slightly lower interest rates. By contrast, also using census survey data to measure the racial composition of a borrower's neighborhood, Kau et al. (2012) reveal that borrowers in predominantly African American neighborhoods or Hispanic neighborhoods tend to prepay less frequently, and after the predicted loan default probability and prepayment probabilities are completely controlled for, only borrowers in predominately African American neighborhoods pay

³⁹ An overage is defined as "a difference between the price at which a loan closes and the minimum price acceptable to the lending institution as quoted on the lender's rate sheet". For details, see Black et al. (2003).

significantly higher contract rates. A recent study by Ghent et al. (2014), utilized a loan data combined with HMDA data to obtain a borrower's race (ethnicity) as well as the racial (ethnic) composition of a borrower's neighborhood, simultaneously investigated the impact of both a borrower's race (ethnicity) and the neighborhood's traits on loan contract rate after the predicted loan default probability is accounted for. Their results provide empirical evidence of adverse pricing against African Americans and Hispanics. The results show both a borrower's race (ethnicity) and the racial (ethnic) composition of a borrower's neighborhood are associated with the loan contract rate after the predicted loan default probability is controlled for. In addition, they also find that the evidence of adverse pricing is strongest for home purchase mortgage loans and loans originated by non-depository institutions.

In terms of methodology, this study is related to Kau et al. (2012) and Ghent et al. (2014). However, Ghent et al. (2014) only consider default risk and do not take prepayment risk into account, while this study models default hazard and prepayment hazard simultaneously through a competing-risks model. Another difference between this study and Ghent et al. (2014) is the current study adopted a discrete-time loan hazard model and successfully addressed the issue of left truncation and right censoring, while Ghent et al. (2014) only modelled default hazard on the loan level, leaving the issue of left truncation and right censoring unsolved. This study differs from Kau et al. (2012) in a way that it employs a matching technique to solve some potential problems with parametric regression analysis. Additionally, Kau et al. (2012) only shed light on the impact of the racial (ethnic) composition of a borrower's neighborhood on loan contract rate, not the effect of a borrower's race (ethnicity) on loan pricing.

3.3. Model

The model in this study is established mainly from a lender's perspective, as a loan's contract rate is determined by the lender at the time of loan origination. Setting the contract rate of a loan requires a lender to meticulously and accurately assess the risk level of the loan at origination, including the likelihood of a borrower to default upon or prepay a loan over the loan's life. In this study, it is assumed that lenders have an accurate model of a borrower's loan termination behavior, and this model is derived from the observation of the actual borrowers' default and prepayment behavior on the loans in the sample of this study. Specifically, the borrower's loan termination behavior is modelled at monthly intervals by observing whether a borrower in the sample defaults upon or prepays a loan each month since loan origination. The lender determines the contract rate of the loan entirely based on those predicted loan termination probabilities.

This analysis framework enables us to investigate whether lenders take a borrower's racial (ethnic) group into account in setting a loan's contract rate beyond the extent to which it affects loan termination probabilities, as the predicted loan termination probabilities are fully accounted for in the lender's loan origination behavior model. The main null hypothesis to test in this study is whether lenders charge the same contract rates for borrowers in different racial and ethnic groups but with the same level of loan termination risk which is represented by the default probability and prepayment probability. Notice that the default risk and prepayment risk are both taken into account, as both default and prepayment would affect the cash flows generated by a loan and its value.

In this study, both a parametric approach and a semi-parametric approach are adopted for analysis. With the use of a parametric approach, both the borrower's loan termination behavior

(loan hazard equation) and the lender's origination behavior (loan contract rate equation) are modelled through parametric regressions using all the loans in the sample. Different from the parametric approach, with the use of a semi-parametric approach, the relationship between a borrower's race (ethnicity) and the contract rate is examined through a matching technique instead of regression, though the borrower's loan termination behavior is modelled in the same way as with the parametric approach so that loan termination probabilities could be predicted. This matching approach is employed based on the hypothesis that borrowers with the same level of loan termination risk should be charged at the same contract rate regardless of their race (ethnicity). Through matching, systematic differences in the level of loan termination risk could be eliminated between two racial (ethnic) groups, therefore a comparison on the contract rate between these two racial (ethnic) groups enables us to investigate whether a borrower's race (ethnicity) itself has an impact on a loan's contract rate. This matching approach, which is aimed to generate a matched sample consisting of comparable loans in terms of loan termination risk across racial groups, and to avoid specifying the loan contract rate equation, is not found in prior racial studies in the mortgage literature.

3.3.1. A Parametric Approach

In order to model the borrower's loan termination behavior, the Cox discrete-time competing-risks model is utilized (Deng et al., 2000), giving the consideration that default and prepayment are believed to compete with each other as a substitute (Kau et al., 1995). One notable feature of this model is a borrower's loan termination behavior is observed and modelled at monthly intervals. This feature is supported given the discrete-time nature of the loan performance data, which is also tracked on a monthly basis. For each month of each loan subsequent to loan

origination, whether a borrower continued, prepaid, or defaulted upon a loan, can simply be identified from the loan performance data. The main purpose of using this discrete-time model is it helps solving the issue of left truncation and right censoring which are fairly common in the mortgage literature.⁴⁰ If the competing-risks model is analyzed only on the loan level by merely classifying loans in the sample into three groups based on whether a loan was continued, or defaulted upon, or prepaid throughout the observation window of loan performance in this study (in other words, only one observation for each loan), the default probability and prepayment probability would be underestimated. In contrast, the discrete-time model, with the assumption that loans when unobservable follow the similar termination pattern as loans when observable, could appropriately estimate loan termination probabilities.

Given the features of this discrete-time competing-risks model, a multinomial logit model at monthly intervals is utilized to model a borrower's loan termination behavior as in Equation (1):⁴¹

$$\ln(p_{jt} / p_{0t}) = \delta_j \alpha_{jt} + \beta_j' \mathbf{x}_{jt} + \varepsilon_{jt} \quad j=1, 2 \quad (1)$$

where p_{0t} is the probability of a loan being continued in period t ; p_{jt} is the probability of a loan being terminated in period t given that this loan had been continued by the beginning of period t , where $j=1$ it is the probability of default at time t , and $j=2$ is the probability of prepayment at time

⁴⁰ Left truncation occurs when a loan had been defaulted or prepaid upon before the start point of the observation window of loan performance, and thus is not observable. Right censoring refers to a situation where a loan has not been defaulted nor prepaid upon by the end of the observation window of loan performance.

⁴¹ Here a widely adopted approach was utilized to estimate this multinomial logit model in which we estimated default hazard and prepay hazard separately, treating other event as censoring and assuming that one event is not informative to the other conditional on all the covariates in this model. This was done mainly because in each hazard model we included almost all of the variables that were likely to affect both of these two events. Another reason is, for loan default and prepayment, there is no theory that could be used as the guideline to impose any parameter restrictions that cross these two hazard equations. Hence, it is not necessary to estimate default and prepay hazard models within a simultaneous equation framework, especially giving the findings that studies show separate models perform well for most of the data (Allison, 2010). In addition, there are other advantages of estimating default and prepay hazard models separately, including the flexibility in specifying different models for different events.

t . In this study, t refers to mortgage time.⁴² The variable α_{jt} is the baseline hazard rate for default and prepayment respectively, and is allowed to vary by mortgage time. The vector of covariates ($\mathbf{x}_{jt}$) describe the observed characteristics of the loan, the borrower, the collateral, the neighborhood, and the economic conditions.⁴³ These covariates may or may not be time varying.

Based on Equation (1), for each loan in each period t , the probability of a loan being defaulted upon or prepaid is predicted. While implementing the predication, it is assumed that among the time-varying covariates in $\mathbf{x}_{jt}$, the term structure, specifically the possible change in the future mortgage interest rate, would be of particular interest to the lender.⁴⁴ In this study, the 10-year treasury constant maturity yield is adopted as the benchmark for mortgage interest rate of 30-year fixed-rate residential mortgage loans. In order to predict the future 10-year yield seen from loan origination, the commonly used Cox, Ingersoll and Ross (CIR) term structure model is employed.⁴⁵ In this term structure model, the whole term structure is assumed to be driven by a spot interest rate ($r(t)$). This spot interest rate is believed to follow a mean-reverting stochastic process with volatility affected by the level of the spot rate as shown in Equation (2):

$$dr(t) = \gamma(\theta - r(t))dt + \sigma\sqrt{r(t)}dz(t) \quad (2)$$

where the first term on the right side of the equation is the deterministic part with θ as the long-term mean of the spot interest rate and γ as the reversion rate, whereas the second part describes the stochastic movements.

⁴² Variables in calendar time could also be expressed with mortgage time, because calendar time could be simply transformed to mortgage time with the use of loan origination month.

⁴³ Covariates on the economic conditions include variables describing the market mortgage interest rate change since loan origination, as interest rate change is a vital financial factor affecting a borrower's decision to prepay and default.

⁴⁴ For other time-varying covariates, we either used the actual values if we were able to observe them, or extrapolated values from the known values, whichever seem more reasonable and appropriate.

⁴⁵ Though several studies in asset pricing argue other interest rate models perform better than CIR term structure model with respect to out-of-sample prediction, those models could only be employed to forecast the mean, not the density of the spot interest rate needed here. In addition, the CIR term structure model is the standard model used in mortgage literature.

Using the estimated parameters in Equation (2),⁴⁶ the density of future spot interest rate for any forecast interval conditional on the spot interest rate at origination $dF(r(t)|r(0))$ is forecasted, based on the transition density of the spot interest rate implied by CIR term structure model.⁴⁷ As the change in the future spot interest rate ($r(t)$) is believed to result in a change in the future 10-year yield ($y_t(r(t))$), and the latter impacts the probability of a borrower defaulting upon or prepaying a loan, the forecasted conditional density of future spot interest rate $dF(r(t)|r(0))$ is employed to predict default probability and prepayment probability for each loan in each period t as in Equations (3) - (5):

$$\hat{p}_{1t}(y_0) = \int p_{1t}[y_t(r(t))] dF(r(t)|r(0)) \quad (3)$$

$$\hat{p}_{2t}(y_0) = \int_{y_t(r(t)) \leq y_0} p_{2t}[y_t(r(t))] dF(r(t)|r(0)) \quad (4)$$

$$\hat{p}_{3t}(y_0) = \int_{y_t(r(t)) > y_0} p_{3t}[y_t(r(t))] dF(r(t)|r(0)) \quad (5)$$

where $\hat{p}_{1t}(y_0)$ is the predicted default probability in period t seen from loan origination given that this loan had been continued by the beginning of period t , $\hat{p}_{2t}(y_0)$ is the predicted pecuniary prepayment probability in period t , and $\hat{p}_{3t}(y_0)$ is the predicted non-pecuniary prepayment probability in period t . Note that, it is assumed that lenders make a distinction between pecuniary

⁴⁶ The parameters in Equation (2) were estimated with the use of 4 time series of yields with different maturities from 1987 to 2007 within the framework of the single-factor CIR term structure model. Those 4 time series are 6-month T-bill yield, 1-year Fama-Bliss bond yield, 3-year Fama-Bliss bond yield, and 5-year Fama-Bliss bond yield. Data were obtained from CRSP. The reason why we chose this estimation period (from 1987 to 2007) is many studies have found there was a shift in Federal Reserve monetary policy in the early 1980s (Duan, Simonato (1999)) and the loan data in this study ends in 2007 based on loan origination year. We used the GAUSS code offered by Jin-Chuan DUAN on his website to implement the estimation, the one used by himself to yield the results in Duan, Simonato (1999). We would like to acknowledge this help from him.

⁴⁷ Notice here, this study forecasts the conditional density of the future spot interest rate rather than the simple conditional mean, as the forecasted density enables the calculation of both predicted pecuniary prepayment probability and non-pecuniary prepayment probability for each loan that are described below, while forecasted mean only allows one to calculate either predicted pecuniary prepayment probability or predicted non-pecuniary prepayment probability for each loan. For the transition density of the spot interest rate, see Cox, Ingersoll, Ross (1985). Here, a normal distribution was used to closely approximate the true transition density.

prepayment and non-pecuniary prepayment. It is deemed that pecuniary prepayment occurs when the 10-year yield at time t (y_t) drops below the yield at origination (y_0), while non-pecuniary prepayment occurs when the future 10-year yield at time t (y_t) is above the one at origination (y_0). The two forms of prepayment are allowed to be differentiated from each other because they are driven by different factors and may have different impacts on the lender. Pecuniary prepayment is normally driven by financial incentives from a dropping interest rate, while non-pecuniary prepayment is probably driven by some non-pecuniary reasons including relocation, divorce, and so forth. Meanwhile, from lenders' perspective, when non-pecuniary prepayment occurs, lenders could reinvest the proceeds from a prepaid loan at a higher interest rate, whereas with the occurrence of pecuniary prepayment, lenders could only reinvest the proceeds at a lower interest rate. Therefore, when predicting loan termination probability at origination, lenders would believe that the occurrence of pecuniary prepayment is more disadvantageous to them than that of non-pecuniary prepayment, and thus require a higher risk premium to compensate the risk of pecuniary prepayment than that of non-pecuniary prepayment. Here, as the forecasted conditional future spot interest rate ($r(t)|r(0)$) is a continuous stochastic variable, in Equations (3) - (5), the integrated expectations are numerically approximated through a discretization approach in which the spot interest rate domain was divided into numerous but finite intervals.

The predicted probability of each termination event in any particular period t , calculated based on Equations (3) - (5), is aggregated over a 10-year span to generate a total predicted probability of each event $\hat{P}_k$ ($k=1, 2, 3$) seen from origination as in Equation (6).⁴⁸ This is in

⁴⁸ In this study, we chose a 10-year span instead of a 30-year span because in reality most of the 30-year fixed rate mortgage loans are prepaid within the first ten years if they were not defaulted upon. Notice here, a capital P is used to distinguish total loan termination probabilities from time-specific loan termination probabilities. The subscript k tells the type of the event, 1 for default, 2 for pecuniary prepayment, and 3 for non-pecuniary prepayment.

accordance with lenders' concern on the total predicted probability of each event over the life of a loan instead of a time-specific predicted probability for a particular period t .

$$\hat{P}_k = \sum_{t=1}^T \left(1 + \frac{y_0}{12}\right)^{-t} \hat{p}_{kt} \prod_{s=1}^{t-1} \left(1 - \sum_{k=1}^3 \hat{p}_{ks}\right) \quad k = 1, 2, 3, \quad (6)$$

Here, $\hat{p}_{kt}$ is the predicted probability of event k in period t given that the loan had been continued by the beginning of period t with the probability as $\prod_{s=1}^{t-1} \left(1 - \sum_{k=1}^3 \hat{p}_{ks}\right)$. Hence, $\hat{p}_{kt} \prod_{s=1}^{t-1} \left(1 - \sum_{k=1}^3 \hat{p}_{ks}\right)$ is the unconditional predicted probability of event k in period t . These probabilities are discounted by the 10-year yield at origination (y_0) with the assumption that loan terminations at earlier stages of the loan are more severe to the lenders. Summation of time-specific discounted unconditional predicted probability of event k over a 10-year window results in a total predicted probability of each event ($\hat{P}_k$).

Loan pricing by lenders mainly depends on the process of approximately assessing a loan's risk level. As the three total predicted probabilities ($\hat{P}_k$) generated from the borrower's termination behavior model properly measure the termination risk level of a loan, within the framework of a parametric approach, they are included in a linear loan contract rate equation (Equation (7)) used to describe how lenders set the contract rate.

$$C_0 = \alpha_0 y_0 + \beta_1 \hat{P}_1 + \beta_2 \hat{P}_2 + \beta_3 \hat{P}_3 + \gamma' \mathbf{z} + \varepsilon \quad (7)$$

In Equation (7), in addition to the three predicted probabilities, some other covariates are also incorporated, including the 10-year yield at origination (y_0) which is believed to be the benchmark used to set the interest rate for 30-year mortgage loans, as well as a set of variables at origination ($\mathbf{z}$) describing the traits of the loan, the borrower, the collateral, and the neighborhood where the collateral is located. A borrower's race (ethnicity) is also included in $\mathbf{z}$, and a linear regression is conducted to test the main null hypothesis in this study: whether a borrower's race

has an additional effect on the contract rate, after the contract rate is rationally determined by the benchmark interest rate and the risk level of the loan.

3.3.2. A Semi-Parametric Approach

Although parametric regression analysis is widely used in empirical studies, especially in racial studies in the mortgage literature, in this study, whether the parametric regression of the loan contract rate determination model (Equation (7)) could result in an unbiased estimation of the effects of a borrower's race (ethnicity) on the contract rate still calls into question, because there might be some potential problems with this parametric regression approach. One potential problem may arise due to the nature of parametric regression technique that the estimation outcomes largely depend on the specifications of the model, specifically the specifications of the contract rate equation (Equation (7)) in this study. If the contract rate equation is mis-specified, the estimation of the race variable in Equation (7) might be biased. In addition, another potential problem in this study is borrowers in different racial (ethnic) groups might have systematically different spectra of loan termination risk. If that is the case, and if the true relationship between the contract rate and a loan risk factor is not a simple linear one and depends on the level of loan risk, a regression analysis based on a single contract rate equation for all of the loans in the pooled sample might fail to provide unbiased estimation. Therefore, the impact of a borrower's race on the contract rate might not be evaluated approximately in such a case. In light of the concerns on these two potential problems, a matching approach is employed, as it does not depend on the specifications of the contract rate equation, and could generate a matched sample of relatively more homogenous loans across racial and ethnic groups.

Matching techniques, aimed to generate a control group of observations comparable to a target group of observations, have been widely used in experimental design and to estimate the causal treatment effects. However, it has been less used to estimate the racial disparities in mortgage contract rate and wage studies⁴⁹ In this study, a matching approach is utilized based on the main null hypothesis (H0) of this study that borrowers having the same level of loan termination risk should be charged at the same contract rate regardless of the borrower's race (ethnicity). Therefore, the main intention of this matching approach is to apply a quasi-experimental design to construct a matched "control" racial group of loans that are as similar as possible to a "target" racial group of loans in terms of loan risk level. Using the matched sample of loans, the impact of a borrower's race (ethnicity) itself on a loan's contract rate could be examined by comparing the average contract rate charged for borrowers in the "target" racial group to the one for borrowers in the matched "control" group to test if there is any systematic difference in contract rate between the two groups, given that any systematic differences in loan termination risk has been mitigated by matching.

Note that in this study, the risk level of a loan is represented by the three predicted loan termination probabilities ($\hat{P}_k$) obtained from the borrower's loan termination behavior model. Therefore, the null hypothesis (H0) in this study could be stated as follows: conditional on the predicted default probability ($\hat{P}_1$), pecuniary prepayment probability ($\hat{P}_2$), non-pecuniary prepayment probability ($\hat{P}_3$), as well as the 10-year yield at origination (y_0), loan contract rate (C_0) should be independent of the borrower's racial group as shown in Equation (8).⁵⁰

⁴⁹ Nopo (2007), Frölich (2007), and Black et al (2008) employed matching technique to examine the impact of gender on wage.

⁵⁰ The 10-year treasury yield at origination (y_0) is also a major factor determining a loan's contract rate, as it is the benchmark of mortgage interest rate for 30-year fixed-rate residential mortgage loans.

$$H0: C_0 \perp Race \mid (\hat{P}_1, \hat{P}_2, \hat{P}_3, y_0) \quad (8)$$

Based on the null hypothesis in Equation (8), nearest 1-to-1 matching between two racial (ethnic) groups is implemented mainly based on the three predicted loan termination probabilities.⁵¹

Specifically, each loan in the “target” racial group is matched to loans in the “control” racial group, selecting the most comparable one as the best match that meets several matching criteria and has the shortest distance on the three predicted loan termination probabilities.⁵² As this is a three-dimensional matching, the Mahalanobis distance that simultaneously takes the three predicted probabilities into account is adopted to measure the distance between the “target” loan and a loan in the “control” racial group.⁵³ This matching approach here thus enables us to construct a matched sample of loans in which loans in these two groups only vary systematically by the borrower’s race (ethnicity), and therefore allows us to investigate whether there is a systematic difference in the contract rate between these two racial groups, the difference that could only be attributed to race (ethnicity) itself.

⁵¹ Notice here, matching in this study is conducted mainly based on the three predicted probabilities generated from the borrower’s loan termination behavior model, not on all of the loan risk factors incorporated to estimate loan termination hazard. The reason why we chose to match based on the predicted termination probabilities instead of the covariates is given the number of loans in the sample and the number of covariates in the loan hazard model, matching based on a high dimensional vector of covariates could barely generate a matched sample with sufficient loans for comparison. In addition, as each of the predicted termination probabilities is a function of those covariates, each probability could be treated as a balancing score. From this point of view, the matching conducted here is similar to propensity score matching that is conducted based on a predicted propensity score (Rosenbaum and Rubin, 1983). However, it is not exactly as same as the typical propensity score matching, as the loans are matched based on the predicted termination probabilities instead of the probability of a borrower being in a particular racial group.

⁵² Several matching criteria are imposed on the maximum difference acceptable for matching in a set of variables between a “target” loan and a loan in the “control” racial group, including the predicted probability of each event, the 10-year yield at origination, as well as several major loan risk factors (e.g., original LTV ratio, FICO score, etc.).

⁵³ Specifically, the Mahalanobis distance (MD) is defined as $MD = (\Delta\hat{P}_1, \Delta\hat{P}_2, \Delta\hat{P}_3) \mathbf{V}^{-1} (\Delta\hat{P}_1, \Delta\hat{P}_2, \Delta\hat{P}_3)^T$, where $\Delta\hat{P}_k$ is the difference in the predicted probability of each event k ($k=1, 2, 3$) between the “target” loan and a loan in the “control” racial group, and $\mathbf{V}$ is the sample covariance matrix of the three predicted termination probabilities from the full set of loans in the “control” racial group. The Mahalanobis distance is preferred because it accounts for the variance of each predicted probability as well as the covariance among the three predicted probabilities. Additionally, the Mahalanobis distance is commonly used for multi-dimensional matching.

3.4. Data

The data used in this study contains 30-year fixed-rate residential mortgage loans serviced by GMAC Residential Capital Company, LLC (GMAC ResCap). GMAC ResCap was a finance firm that provided home financing, loan serving, and mortgage-back securities (MBS) issuance in the U.S. by the recent financial crisis. Loans in this data were all packaged into private-label mortgage-backed securities and traded in the secondary mortgage market. As the loan servicer, GMAC ResCap collected detailed information on the loan, the borrower, and the collateral at the time of loan origination. In addition, it also tracked the performance of each loan it serviced on a monthly basis. The monthly loan performance data provides information on the current interest rate and balance of a loan, as well as prepayment and delinquency status. The monthly loan performance data and the loan origination data, both from GMAC ResCap records, were matched together through a unique loan identification number created by this servicer. In this study, default is defined as the occurrence of a borrower being 90-days delinquent, and that occurrence eventually results in a foreclosure.

Though the loan origination records provide detailed information on a borrower's financial characteristics, they do not have any information on a borrower's race. However, information on a borrower's race (ethnicity) group is available in the Home Mortgage Disclosure Act Loan/Application Register (HMDA-LAR) data.⁵⁴ The Home Mortgage Disclosure Act (HMDA) is a federal law that was enacted by the Congress in 1975 and implemented by the Federal Financial

⁵⁴ Ethnicity is new filed added to the HMDA LAR data beginning January 1st, 2004. In 2002, The Federal Reserve Board amended Regulation C's rules for lending institutions to collect information on an applicant's race and ethnicity, and the revised HMDA requirements started to apply to loan applicants on January 1st, 2004. Before 2004, an applicant was required to only self-identify his/her race, from among the five choices available (American Indian or Alaskan Native, Asian or Pacific Islander, African American, Hispanic or Latino, and White). Since 2004, an applicant is required to self-identify whether he/she is a Hispanic or Latino under the category of ethnicity. In addition, he/she is also required to identify a race, from among the new five choices available (American Indian or Alaskan Native, Asian, African American, Native Hawaiian or Pacific Islander, and White).

Institutions Examination Council (FFIEC). Starting at 1989, on the basis of the amendments to HDMA resulting from the Financial Institutions Reform, Recovery, and Enforcement Act (FIRREA), financial institutions are required to collect and disclose data on race (ethnicity), gender, and income of the applicant and the co-applicant (if applicable) for each loan application (the LAR data) instead of the aggregate data on the census tract level they reported before. Thus the loan-level HMDA-LAR data since 1989 contains detailed information on the personal traits of an applicant (co-applicant). In addition, one notable advantage of the HMDA-LAR data is a great number of lending institutions are covered by HMDA.⁵⁵ It is estimated that in 2007, just before the start of the recent financial crisis, more than 8,800 lending institutions reported their data to FFIEC, accounting for approximately 80% of all home lending nationwide (Avery et al., 2007).

In order to make the use of information on a borrower's race (ethnicity), the loan data from GMAC ResCap was matched with the HMDA-LAR data. However, one issue arises – these two data sets define the location of collateral differently. Loans in the GMAC ResCap data are reported on the zip code of the underlying property, while loan applications in the HMDA-LAR data are reported on the basis of the census tract, and the definition of a census tract adopted in HMDA-LAR data varies across time.⁵⁶ To address this issue, we first matched the GMAC ResCap loan data to a source of readily available property transactions data to identify the geographic coordinates of the property securing each loan. Hence, the 1980, 1990, and 2000 census tract identification numbers for each loan could be identified and utilized to match loans with approved loan applications in the HMDA-LAR data.

⁵⁵ The FIRREA amendments in 1989 expanded the coverage of HMDA to many independent non-depository lending institutions, in addition to the previously covered savings associations, banks, credit unions. For detailed information on who are required to report HMDA data, see the descriptions on Federal Financial Institutions Examination Council (FFIEC) website <https://www.ffiec.gov/hmda/reporter.htm>.

⁵⁶ HMDA LAR data from 1989 to 1991 used the 1980 census tract boundaries, data from 1992 to 2002 followed the 1990 census tract definitions, and data from 2003 to 2012 adopted 2000 census tract definitions.

The GMAC ResCap loan data in this study were restricted to 30-year first-lien fixed-rate residential home purchase mortgage loans with underlying properties located in Miami-Dade County, FL to provide for a uniform sample, and to match with the properties transaction data in Miami-Dade County which is publicly available.⁵⁷ This loan-property sale matching is conducted based on several common fields including the value of the underlying property, property sale month (loan origination month), property type, and zip code, leading to a matched sample of loan-property sales.⁵⁸ Each matched loan-property sale with the identified 1980, 1990, and 2000 census tract numbers was then matched with conventional loan applications that were approved and originated in Miami-Dade County, FL in the HMDA-LAR data based on another set of common fields. Those common fields are loan amount (in thousand dollars), loan origination year (loan action year), property type, property occupancy status, loan purpose, lien status (if applicable), and census tract.⁵⁹ Only loans with a unique HMDA-LAR match are included for analysis in this study.⁶⁰

One issue with the use of a borrower's race (ethnicity) information in the HMDA-LAR data is the reporting rules changed significantly with regard to the classification of race and ethnicity in 2004. Prior to 2004, an applicant (co-applicant) was required to self-identify his/her race from among the five racial classifications – American Indian, Asian, African American,

⁵⁷ The property transaction data is provided by the Office of the Property Appraiser in Miami-Dade County, FL.

⁵⁸ Each mortgage loan was matched to property sales in the pool with replacements requiring that the gap between the appraised value of the property in the loan data and the transaction price of the property in property sale data is the minimum one in the pool. If multiple property sales were matched to a loan and those multiple properties are located in the same census tract, this loan is treated as having a unique property sale match, and its census tract identification number is identified as the common one for those multiple properties.

⁵⁹ As the loan amount in HMDA-LAR data is in thousand dollars, we allowed loan amounts to differ by up to \$1,000. Lien status is a new field added to the HMDA LAR data beginning January 1st, 2004.

⁶⁰ In this study, if there are multiple HMDA-LAR matches for a loan that meet those matching criteria above, if those multiple matches have exactly the same race information on the borrower and the co-borrower, this loan is allowed to be identified as having a unique HMDA-LAR match, as the goal of the match is to obtain the race information on the borrower(s).

Hispanic, and White. However, since 2004, applicants were asked to separately report their ethnicity (Hispanic or non-Hispanic) and race (American Indian, Asian, African American, Native Hawaiian, and White). In order to align the race and ethnicity data since 2004 with the data prior to 2004, a hierarchy among all the race and ethnicity classifications is adopted in this study.⁶¹ Based on this hierarchy, the three race groups – African American, American Indian, and Asian – “trump” Hispanic, and Hispanic “trumps” White. Specifically, if an applicant since 2004 self-identified as an African American/American Indian/Asian, no matter whether he/she identified as a Hispanic or not, he/she is treated as an African American/American Indian/Asian. If an applicant reported Hispanic for the ethnicity and White for the race, he/she is treated as a Hispanic. Therefore, White borrowers could only be non-Hispanic Whites. In addition, as both the applicant and the co-applicant (if applicable) are required to report their race (ethnicity) group, another issue with the HMDA-LAR data is how to represent the race (ethnicity) group of a loan application if the applicant and the co-applicant self-identified as in different racial (ethnic) groups. The hierarchy described above is also used to solve this problem with a reasonable ordering: African American, American Indian, Asian, Hispanic, and finally non-Hispanic White.⁶²

The matched loan-property-HMDA data were supplemented with other data sources used to describe the characteristics of the neighborhood where the underlying property of a loan is located, based on the property’s location. In this study, a property’s neighborhood is defined on the basis of the 1990 census tract boundaries. The decennial census survey data was incorporated to generate time-varying variables describing a neighborhood’s traits (including housing

⁶¹ Avery et al. (2007) discussed several hierarchies to solve this issue, and this study adopted one hierarchy that is reasonable here given the demographic characteristics of the population in Miami-Dade County, FL.

⁶² For example, if either the applicant or co-applicant chooses African American as one of their races, the loan application is classified as a loan by an African American. Another example is if one applicant chooses Hispanic, and the other applicant chooses non-Hispanic White, the loan application is classified as a loan by a Hispanic.

occupancy rate and poverty rate) using a linear time-trend between the decennial data from 1990 to 2000 and from 2000 to 2010, normalized to the 1990 census tract boundaries. The HMDA data aggregated on the census tract level was utilized to create measures of the demographic characteristics of each census tract on a yearly basis.⁶³ Specifically, those measures, including the median applicant income and the racial (ethnic) composition of a neighborhood, were calculated for each census tract following the 1990 census tract boundaries, and for each year based on a three-year window (the previous year, the current year, and the next year), using all of the loan applications including applications either approved or denied for 1-to-4 family dwelling home purchase in a census tract across all lenders.⁶⁴ The resulting yearly census-tract level measures were further linearly time-trended to generate monthly measures.

Changes in house price and the variation in housing sale prices within a neighborhood (census tract) were measured using the property transactions data.⁶⁵ A median housing price index was generated for each census tract based on the 1990 census tract definitions, and for each month of analysis by creating a three-year window of sale, eighteen months before and eighteen months after, calculating the inflation-adjusted median price.⁶⁶ The resulting index is a unique median house price index for each census tract on a monthly basis. This index is used to measure the changes in house prices over the life of the loan relative to loan origination. The standard deviation

⁶³ In this study, the census-tract level aggregated HMDA data is preferred rather than the decennial census survey data to generate time-varying variables on the demographic characteristics of each census, because the aggregated HMDA data which is updated every year tends to be more accurate than census survey data which is updated every ten years.

⁶⁴ Those measures include the median applicant income, the proportion of African American applicants, the proportion of non-Hispanic White applicants, and the proportion of Hispanic applicants. The applicant income is inflation adjusted by a GDP per capita deflator. All income is defined in 2009 dollars. The racial and ethnical group of a loan application is identified using the hierarchy described above. The calculation is conducted also based on the 1990 census tract boundaries.

⁶⁵ The data on the pool of property transactions are sales from 1990 to 2013 in Miami-Dade County, FL.

⁶⁶ The median housing sale price is preferred to an average housing sale price in order to prevent any extreme housing sale prices in a census tract from affecting the measurement of the overall house price level. The housing sale price is inflation adjusted by a GDP per capita deflator. All prices are defined in 2009 dollars.

of the housing sale price is also calculated monthly for each census tract using a three-year window. It is utilized to measure the heterogeneity in housing sale price in a neighborhood. Additionally, for each loan at loan origination, the neighborhood-level recent housing price appreciation rate is calculated using the growth rate of the calculated median housing sale price within a pre-origination window.⁶⁷

There were initially 1,994 loans originated in Miami-Dade County, FL that could be matched to the transaction and HMDA-LAR data sets. Loans with missing values on the loan, the borrower, the collateral or the neighborhood characteristics, or loans without completed loan performance data were deleted. The sample is restricted to loans originated from 1997 to 2006, because very few loans were originated outside this study window. In addition, because only a few loans were originated to American Indian or Asian borrowers, they were removed from the data. The final sample consists of 1,404 observations of 30-year first-lien fixed-rate residential mortgage loans for home purchase, and they are all subprime loans given the fact that GMAC Rescap specialized in servicing subprime residential mortgage loans. Loans in the final sample were originated from Jan. 1997 to Dec. 2006. Their performance was tracked from Jan. 2000 to Oct. 2010, a period covering the recent mortgage crisis.

Table 3.1 presents a brief description of the characteristics of the loans at the time of loan origination in the pooled sample, including the traits of the loan, the borrower, the collateral, and the neighborhood. The feature of high credit risk of these loans is reflected in the descriptive statistics of the variables including the original LTV, the proportion of borrowers lacking full income documentation, and the proportion of loans encumbered by a prepayment penalty. The

⁶⁷ Recent housing price appreciation rate at origination is defined as the ratio of the median housing sale price in a neighborhood in a three-year period prior to the month of loan origination to the median housing sale price in the same neighborhood in another three-year period prior to the three-year pre-origination period, then minus 1.

average original LTV ratio in this sample is around 85%, and 88 loans (6.27%) have a LTV ratio greater than 100%, indicating borrowers of the 88 loans did not have any equity on their house at loan origination. In addition, more than half of the borrowers (52.35%) failed to provide full income documentation, and 18.52% of the borrowers (260 loans) have a FICO score below 650. Only 67.02% of the loans are not encumbered by a prepayment penalty giving the fact that a prepayment penalty is fairly rare for residential mortgage loans. Not surprisingly, the average contract rate charged for the loans in the sample is relatively high. As shown in Table 3.1, the average contract rate is 7.96%, approximately 304 basis points higher than the 10-year treasury yield at origination. During the observation period of loan performance in this study, of the 1,404 loans, 122 loans (8.69%) were defaulted upon and ended up with a foreclosure, while 1,129 loans (80.41%) were prepaid.

Table 3.2 provides descriptive statistics of the loans by the racial (ethnic) group of a borrower(s). Around 70.87% of the loans (995 loans) in the sample were by Hispanic borrowers, about 22.01% of them (309 loans) were by non-Hispanic White borrowers, and approximately 7.12% of the loans (100 loans) were by African Americans. Those proportion numbers are close to the proportion numbers from the census survey describing the racial (ethnic) composition of all the residents in Miami-Dade County, FL. This indicates that the sample in this study well represents the overall Miami-Dade population.⁶⁸

As indicated in Table 3.2, borrowers of different race (ethnicity) appear to have different loan termination patterns. The observed average default rate within the study period by African

⁶⁸ The proportion of African American borrowers in the sample of this study is slightly lower than the proportion of African American residents in Miami-Dade County, FL by the census survey (approximately 20% in 2000 census survey data). This leaves a research question for future study on whether African Americans face more difficulties having access to credit than borrowers in other racial and ethnical groups. As we do not have data on loan applications, this research question is beyond the scope of this study.

Americans is around 13.00%, approximately twice as high as the observed default rate by non-Hispanic Whites (6.47%). Meanwhile, the observed prepayment rate by African Americans (77.00%) or by Hispanics (78.39%) is lower than that by non-Hispanic Whites (88.03%). The observed differences in loan termination patterns among the three groups might be explained by the differences in the credit risk of a borrower. As shown in Table 3.2, African Americans appear to have the highest credit risk simply based on the average original LTV and the FICO score, while non-Hispanic Whites appear to have the lowest risk. Additionally, African Americans seem to be more likely to provide full income documentation at loan origination, and more likely to be encumbered by a prepayment penalty. The descriptive statistics also seem to indicate that African American borrowers tend to live in a neighborhood with a relatively lower recent housing price appreciation rate, a higher poverty rate, and lower median income. The observed differences in the characteristics of the loan, the borrower, and the neighborhood among the three racial (ethnic) groups indicate the importance of accounting for all of these traits when examining the relationship between a borrower's race (ethnicity) and loan hazard probability/ loan contract rate.

3.5. Model Specifications

3.5.1. Regression Specifications

In the discrete-time competing-risks model (Equation (1)), the baseline hazard rate (α_{jt}) and a set of covariates ($\mathbf{x}_{jt}$) are incorporated to model a borrower's loan termination behavior in each period t . On the basis of the option-theoretic model of mortgage loan "financial" termination (Kau et al., 1995), default and prepayment are driven by different financial incentives. Therefore, different baseline hazard rates (α_{jt}) are applied for default and prepayment, respectively. While a

scaled standard Default Assumption schedule (SDA) is used for default hazard, mortgage year fixed effects are employed to measure prepayment baseline hazard rate.⁶⁹

Also based on the option-theoretic model of “financial” termination (Kau et al., 1995), market interest rate change and change in the value of the collateral are the two most prominent time-varying factors affecting a borrower’s decision to default or prepay. In this discrete-time model, the market interest rate change at time t is measured by the gap between the 10-year treasury constant maturity yield at loan origination and the 10-year yield at time t lagged by 2 periods ($y_0 - y_{t-2}$).⁷⁰ To measure the change in the value of the underlying house, a monthly census tract-level median housing price index is employed which is generated by the pool of residential property transactions in Miami-Dade County, FL, with the assumption that the value of a house changes at the same pace as the median housing price index in the census tract where the house is located. For each mortgage time t , in order to reflect the extent to which the house price changes at time t relative to the house price at loan origination, the ratio of the median house price index level at time t to the one at origination is calculated, and this ratio is named as relative house price at time t (RHP_t). Additionally, an interaction term of the market interest rate change ($y_0 - y_{t-2}$) and relative house price at time t (RHP_t) is also included in this competing-risks model to account from any correlation between interest rate change and house price change.

Variables describing the traits of the loan, the borrower, the property, and the neighborhood are also included in the vector of $\mathbf{x}_{jt}$. They are the loan contract rate spread at origination ($C_0 -$

⁶⁹ The traditional Public Securities (PSA) schedule is not used because previous studies argued this schedule did not describe the pattern of actual prepayments well, for more details, see Kau et al. (2004).

⁷⁰ Notice that, the 10-year treasury constant maturity yield at time t lagged by 2 periods is used as the yield at time t for every mortgage time t except for the first mortgage month and the second mortgage month. For these two months, the 10-year yield at loan origination (at time 0) is used. A yield at time t lagged by 2 periods is used because in practice there is usually a gap between a borrower’s decision and actual termination, and borrowers typically rely on past information to make their decisions.

y_0), original LTV, FICO score of the borrower, original loan amount, loan origination season fixed effects, and dichotomous (0,1) indicators for whether a borrower provided full income documentation at origination, whether the loan is encumbered by a prepayment penalty at time t , whether the underlying property is occupied by the owner, and whether the property is a single family detached house or a condo.⁷¹ Also, several time-varying variables are included to measure the characteristics and evolution of the neighborhood of the underlying property, including the heterogeneity in the housing sale price, housing occupancy rate, poverty rate, and the median income of loan applicants at time t .⁷² The individual- and neighborhood-level race (ethnicity) variables at time t are included in this model to test if they are associated with loan default hazard or prepayment hazard after the set of loan risk factors are accounted for.⁷³ Note here, in the default equation, prepayment penalty variable at time t is excluded, whereas in the prepayment equation, the original LTV ratio, the income documentation status variable, and the variable of neighborhood-level heterogeneity in housing price at time t are excluded, because these variables are not believed to directly affect the corresponding hazard respectively.⁷⁴

In this discrete-time competing-risks model, the loan contract rate spread at origination ($C_0 - y_0$) might be endogenous. If the contract rate (C_0) is endogenous, this variable is endogenous

⁷¹ Contract rate spread at origination is defined as the gap between the contract rate and the 10-year yield at origination. We chose contract rate spread instead of contract rate itself because the 10-year yield as the benchmark mortgage interest rate varied considerably within our study period, and this spread allows us to make comparison within mortgage vintage.

⁷² Heterogeneity in housing sale price at time t was measured by the standard deviation of housing sale price within a three-year window prior to time t in a neighborhood. Housing occupancy rate and poverty rate in a neighborhood at time t are from the census survey data. Median income of loan applicants in a neighborhood at time t is generated by the aggregated HMDA data described in the data section. In order to reflect the rank of each census tract in terms of median applicant income in Miami-Dade County, FL at time t , the ratio of the median applicant income in a census tract at time t to the median applicant income in Miami-Dade County, FL at time t is calculated and used here.

⁷³ The variables describing the racial composition of loan applicants in a census tract at time t are generated by the aggregated HMDA data described in the data section. Notice here, those neighborhood-level race (ethnicity) variables are only incorporated in specification (2) in Table 3.3 and 3.4.

⁷⁴ We also tested whether those excluded variables affect the corresponding hazard respectively. None of the estimated coefficients of those variables are statistically significant.

simply because it is a linear function of the contract rate. The contract rate might be endogenous because as a lender believes a borrower is more likely to default or prepay, the lender would charge a higher contract rate, meanwhile being charged at a higher contract rate, a borrower might be more likely to terminate a loan. The control function (CF) method is utilized to solve the issue, as this competing-risks model specified as a multinomial logit model is not a linear one. Using the CF method, a reduced-form estimation of the contract rate is implemented in which the contract rate is regressed against all of the exogenous variables in the system, and the residual from this estimation is included in the multinomial logit model to solve the potential issue of endogeneity.

Turning our attention to the contract rate determination equation (Equation (7)) used to describe a lender's origination behavior, the issue of endogeneity also draws our attention that the three predicted loan termination probabilities ($\hat{P}_k$) are endogenous. Recall that the three predicted probabilities are generated from the competing-risks model in which the contract rate spread ($C_0 - y_0$) is included. Therefore, each of the three predicted probabilities ($\hat{P}_k$) is a function of the contract rate, and is endogenous in the final contract rate determination equation. A set of generated IVs ($\tilde{P}_k$) are used to solve this issue, each as a valid IV for the corresponding generated variable ($\hat{P}_k$). Each generated IV ($\tilde{P}_k$) is created following the same procedure to calculate its corresponding generated variable ($\hat{P}_k$), but with the actual contract rate spread ($C_0 - y_0$) being replaced with a predicted contract rate spread ($\hat{C}_0 - y_0$). The predicted contract rate spread is the difference between a predicted contract rate at origination ($\hat{C}_0$) and the actual 10-year yield at origination (y_0). The predicted contract rate ($\hat{C}_0$) is obtained by the reduced-form contract rate equation estimation used in the CF method. As the predicted contract rate ($\hat{C}_0$) is a function of all the exogenous variables, each generated IV ($\tilde{P}_k$) is hence exogenous and serves as a valid IV for its corresponding

generated variable ($\hat{P}_k$) in the loan contract rate equation. With those generated IVs, 2SLS is used to estimate the contract rate determination model, since this model is a linear one.

The loan contract rate determination model (Equation (7)) also includes the 10-year yield at origination (y_0), as well as a set of covariates included in $\mathbf{z}$ that theory suggests should affect loan pricing. These additional covariates include the original LTV, FICO score of the borrower, original loan amount at origination, loan origination season fixed effects, prepayment penalty fixed effects, a dichotomous variable for whether a borrower provided full income documentation at loan origination, underlying property type fixed effects, property occupancy status fixed effects, and a list of neighborhood characteristics at loan origination including recent house price appreciation rate, heterogeneity in housing sale price, housing occupancy rate, poverty rate, and median income of loan applicants.⁷⁵ Variables on the racial composition of loan applicants in a census tract at origination are also incorporated to investigate the issue of redlining.⁷⁶ Additionally, a trend term in calendar time was also included. Table B.1 in the Appendices lists all of the variables used in this study and provides a detailed description of each variable.

In both the loan hazard model and loan contract rate determination model, some variables take a nonlinear function form based on prior studies or some theoretical reasons. The original LTV ratio is transformed into categorical variables: loans with LTV ratio less than or equal to

⁷⁵ Recent house price appreciation rate at origination in a neighborhood was described by the growth rate of the median housing sale price prior to loan origination. Specifically, it was defined as the ratio of the median housing sale price in a neighborhood in a three-year period prior to month of loan origination to the median housing sale price in the same neighborhood in another three-year period prior to the three-year pre-origination period just mentioned, then minus 1. Heterogeneity in housing sale price at origination was defined as the standard deviation of the housing sale price over a three-year period prior to the month of loan origination. Median income of loan applicants in a census tract at origination is generated by the aggregated HMDA data elaborated in the data section. Here, in order to reflect the rank of each census tract in terms of median applicant income in Miami-Dade County, FL at origination, the ratio of the median applicant income in a census tract at loan origination to the median applicant income in Miami-Dade County, FL at origination is calculated and used here.

⁷⁶ Variables on the racial composition of loan applicants in a census tract at origination are generated by the aggregated HMDA data described in the data section. As shown in Table 3.5, they are only included in specification (2).

80%, loans with LTV ratio greater than 80% but less than or equal to 90%, loans with LTV ratio greater than 90% but less than or equal to 100%, and loans with LTV ratio exceeding 100%. Furthermore, FICO score is entered as a continuous linear spline function with a knot point at 700 based on the assumption that once the FICO score is above 700, an additional increase in FICO score does not have any effects on loan termination probabilities/loan contract rate.⁷⁷ Finally, following prior studies, a quadratic function form of the original loan amount is also utilized to allow for a non-linear relationship.

3.5.2. Matching Algorithm

The Matching technique is applied to mitigate any systematic differences in the level of loan termination risk by matching loans in a particular racial (ethnic) group to those in another group based on the three predicted probabilities ($\hat{P}_k$). For each loan in the “target” racial (ethnic) group in the sample, a nearest 1-to-1 matching with replacements and with calipers is performed to find a loan in the “control” racial (ethnic) group which is as close as possible to the “target” one. The matching is implemented between loans by African American borrowers and loans by non-Hispanic White borrowers, and between loans by Hispanic borrowers and non-Hispanic White borrowers, respectively.⁷⁸

⁷⁷ The FICO linear spline function was specified as follows: $FICO_{(FICO \leq 700)} = \text{minimum}(FICO, 700)$; and $FICO_{(FICO > 700)} = \text{maximum}(FICO, 700) - 700$. Therefore, coefficient on $FICO_{(FICO \leq 700)}$ measures the effects of FICO score on dependent variable when $FICO \leq 700$; while coefficient on $FICO_{(FICO > 700)}$ measures the effects of FICO score when $FICO > 700$. We tested whether the results are robust to the specification of the FICO score knot point by conducting the same analysis with a knot point at 720 or 750, and results are robust.

⁷⁸ When matching loans by African American borrowers with loans by non-Hispanic White borrowers, the “target” racial (ethnic) group is the group of African American borrowers, whereas the “control” group is the group of non-Hispanic White borrowers. For matching between loans by Hispanic borrowers and loans by non-Hispanic White borrowers, the “target” group is the group of non-Hispanic White borrowers, whereas the “control” group is the group of Hispanic borrowers, because in the pooled sample of this study, the number of loans by non-Hispanic Whites is far less than that by Hispanic borrowers.

The matching algorithm is described as follows. First, as the matching is conducted mainly on the three predicted loan termination probabilities, before matching the density distribution of each predicted probability ($\hat{P}_k$) in the “target” racial (ethnic) group is compared to that of the “control” group to check if they overlap with each other. Loans in the “target” group that do not fall in the region of common support are not included for matching. Next, each loan in the “target” racial (ethnic) group within the common support is matched to all of the loans in the “control” group in the sample, but only those loans that satisfy a set of matching criteria named calipers are selected as potential “matches”. A caliper is applied to impose a tolerance level for the maximum distance on a matching criterion variable between borrowers in the two racial (ethnic) groups. Those matching criterion variables include the predicted probability of each event ($\hat{P}_k$), the 10-year yield at origination (y_0), as well as several variables that are believed to be the major loan risk factors including the original LTV ratio, FICO score, and original loan amount. Calipers are imposed in order to largely reduce the bias in terms of the loan risk between the two racial (ethnic) groups. As the caliper radius (the width of a caliper) affects the performance of the matching technique, specifications of the caliper radius is of great important. Following most of the prior studies using matching techniques, the caliper radius on a matching criterion variable is specified as a ratio (ρ) of the pooled standard deviation of that variable.⁷⁹ Here, a uniform caliper radius ratio (ρ) is applied to all of the matching criterion variables listed above.⁸⁰ Then with the resulting

⁷⁹ Specifically, the caliper is defined as $Caliper = \pm \rho \sqrt{(s_{group1}^2 + s_{group2}^2) / 2}$, where s_{group1}^2 and s_{group2}^2 denote the sample variance of a matching criterion variable in racial (ethnic) group 1 and group 2, respectively, whereas ρ is the ratio. Previous studies (Cochran and Rubin, 1973; Rosenbaum and Rubin, 1985; Austin, 2011) examined the extent to which each particular ratio (ρ) ranging from 0.05 to 2.50 could reduce the bias between the two groups. In this study, different values of this ratio (ρ) is used to find the optimal one that could result in a balanced matching sample with sufficient loans.

⁸⁰ This caliper ratio is applied to all the matching criterion variables described above except for the original LTV ratio. For the variable of original LTV ratio, it is required that those potential “matches” should fall in the same original LTV ratio category as the “target” loan. Loans in the sample are classified into 3 groups based on the original LTV

set of potential “matches” from the first step, the loan with the smallest Mahalanobis distance on the three predicted loan termination probabilities ($\hat{P}_k$) is selected as the sole and best match, and is included in the matched “control” racial (ethnic) group.⁸¹

Matching steps described above thus generate a matched sample of loans, and its balance needs to be diagnosed in order to ensure that the two racial (ethnic) groups of borrowers have the same distributions of loan termination risk and observed traits on the loan, the borrower, the collateral, and the neighborhood.⁸² The standardized difference of the mean, the ratio of the variance, and mean difference *t*-test of all the variables in Equation (7) between the two racial (ethnic) groups are used for matching balance diagnosis following the matching literature (Rubin, 2001; Austin, 2011).⁸³ Because whether the matched loan sample is balanced or not may depend on the caliper radius (the ratio (ρ)), if a matched sample with a specific caliper radius is not balanced, the matching steps illustrated above are repeated with other values of caliper radius until

ratio: loans with original LTV ratio below 80, loans with original LTV ratio above 80 but below 100, and loans with original LTV ratio above 100.

⁸¹ Specifically, the Mahalanobis distance (MD) is defined as $MD = (\Delta\hat{P}_1, \Delta\hat{P}_2, \Delta\hat{P}_3) \mathbf{V}^{-1} (\Delta\hat{P}_1, \Delta\hat{P}_2, \Delta\hat{P}_3)^T$, where $\Delta\hat{P}_k$ is the difference in the predicted probability of each event k ($k=1, 2, 3$) between the “target” loan and a loan in the “control” racial group, and $\mathbf{V}$ is the sample covariance matrix of the three predicted termination probabilities from the full set of loans in the “control” racial group.

⁸² The balance is checked not only on the three predicted loan termination probabilities that are believed to measure the risk level of a loan, but also on the covariates incorporated in the loan hazard model (Equation (1)) and in the loan contract rate determination equation (Equation (7)), as those covariates describe the characteristics of the loan, the borrower, the collateral, and the neighborhood. The main reason for it is to form a matched sample of loans in which the only difference between the two racial (ethnic) groups is the borrower’s race (ethnicity).

⁸³ Given the fact the mean difference *t* test is criticized by many authors (Imai et al., 2008; Austin, 2008b, 2009b), because the significance levels are confounded with sample size, the standardized difference of the mean and the ratio of the variance of each variable are used to check if the matched sample is well balanced or not. Specifically, for a single continuous variable x , the standardized difference of the mean is defined as $SD = (\bar{x}_{group1} - \bar{x}_{group2}) / \sqrt{(s_{group1}^2 + s_{group2}^2) / 2}$, where $\bar{x}_{group1}$ and $\bar{x}_{group2}$ denote the sample mean of this variable in racial (ethnic) group 1 and group 2, respectively, whereas s_{group1}^2 and s_{group2}^2 denote the sample variance of the variable in racial (ethnic) group 1 and group 2, respectively. For a dichotomous variable, the standardized difference of the mean is defined as $SD = (\bar{x}_{group1} - \bar{x}_{group2}) / \sqrt{(\bar{x}_{group1} \times (1 - \bar{x}_{group1}) + \bar{x}_{group2} \times (1 - \bar{x}_{group2})) / 2}$. Following Rubin (2001), a variable is well balanced if and only if the standardized difference of the mean falls in the range of (-0.25, 0.25), and the ratio of the variance falls in the range of (0.5, 2).

a balanced matching sample is arrived at in which there are sufficient loans for comparison across the two racial (ethnic) groups.

Using the balanced matching sample, the average contract rate charged on loans in the “target” racial (ethnic) group is compared to that in the “control” group to test if there is any significant difference between the two groups, the difference that is not attributed to differences in the risk level of the loan or the observed traits, but just the borrower’s race (ethnicity). Additionally, a regression adjustment as specified in Equation (7) is also applied to each matched sample of loans to examine if a borrower’s race (ethnicity) has an additional impact on a loan’s contract rate.⁸⁴

3.6. Results

3.6.1. Regression Results

Table 3.3 reports the estimation results on loan default hazard, as part of the discrete-time competing-risks model. Model (1) includes a borrower’s race (ethnicity) in the estimation of the likelihood of a loan being defaulted upon. Model (2) includes a borrower’s race (ethnicity) as well as the neighborhood-level race (ethnicity) variables. Both models explain the impact of race (ethnicity) on loan default probability.

As shown in Table 3.3, a borrower’s decision to default is mainly driven by financial incentives. The change in the value of the collateral is one of the most prominent variables affecting default probability, as the coefficient on the variable of relative house price at time t (RHP_t) is statistically significant and negative. This is consistent with the option theory of

⁸⁴ The regression adjustment is applied to a matched sample in the scenario where the matched sample is balanced with regard to most of the variables, especially the predicted default probability, prepayment probabilities, and variables on the characteristics of the loan, the borrower, and the collateral, but a few less important variables are not well balanced.

mortgage loan termination (Kau et al. 1995) – as house price decreases, borrowers losing equity on their houses, tend to be more likely to default. In addition, estimation results on the variables describing the traits of the loan, the borrower, and the collateral, demonstrate that borrowers with an original LTV ratio above 100%, a lower FICO score, a larger original loan size, limited income documentation, or borrowers using a loan to purchase a house as investment rather than as their primary residence tend to be more likely to default on the loan. Note that the estimation results on the two FICO score variables indicate that when a borrower's FICO score is above a threshold (e.g.: 700), an additional increase in the FICO score does not have any effects on the likelihood of a borrowing defaulting upon a loan. Meanwhile, the impact of the original loan size on default probability lessens as the original loan size increases. Among the neighborhood-level variables describing the characteristics of the local housing market and the economic conditions, only the estimate on the variable of median applicant income is significant and negative. The results indicate that borrowers living in a relatively poorer neighborhood tend to be more likely to default. All of the results are consistent with most of the prior studies.

The results on loan default hazard, reported in Table 3.3, do not provide any empirical evidence that a borrower's race (ethnicity) is associated with loan default probability, as the estimates on the individual-level race (ethnicity) variables are not significant in either Model (1) or Model (2). This finding indicates that the differences in the observed default rate across various racial (ethnic) groups as reported in Table 3.2 could be entirely explained by the differences in the covariates incorporated in this default hazard model. In addition, the neighborhood-level race (ethnicity) variables are not shown to affect loan default probability either.

Table 3.4 demonstrates the estimation results on prepayment hazard. Similar to Table 3.3, in Table 3.4, neighborhood-level race (ethnicity) variables are excluded in Model (1), but are

included in Model (2). Overall, the results in these two models are consistent, and are in accordance with theoretic expectations and prior studies. As shown in Table 3.4, the market interest rate change plays an important role in a borrower's decision to prepay. As the market interest rate drops, borrowers tend to be more likely to prepay. This is consistent with the option theory of mortgage loan termination (Kau et al. 1995), as rational borrowers always have financial incentives to take the advantage of a lower interest rate through prepayment. In addition, the likelihood of a borrower prepaying a loan is shown to be also affected by the change in the value of the collateral. In a market seeing great house price appreciation, borrowers would have more incentives to prepay their loans to extract the equity on their house. Besides, another explanation for the significantly positive relationship between prepayment probability and relative house price at time t (RHP_t) is default and prepayment are believed to compete with each other, as house price increases, borrowers are less likely to default, and hence more likely to prepay.

In addition to market interest rate change and house price change, other factors are shown to affect loan prepayment probability also. The contract rate spread at origination ($C_0 - y_0$) appears to be positively associated with loan prepayment probability. The contract rate spread ($C_0 - y_0$) measures the extent to which the contract rate at closing (C_0) deviates from the baseline interest rate (y_0), and thus risk premium. The results here indicate that a borrower with a higher risk premium tends to be more likely to prepay. Another more plausible explanation for this result is the effects of the points used to buy down a contract rate. As theory suggests, a borrower, anticipating a lower likelihood of prepaying a loan in the near future, self-selects more points at loan closing in exchange for a reduced contract rate. From this point of view, the likelihood of a loan being prepaid is anticipated to be positively associated with loan contract rate, thus the contract rate spread. In this study, due to the lack of data on points, this possible explanation could

not be examined. Estimation results on other covariates demonstrate that borrowers with a higher FICO score or a larger loan size appear to be more likely to prepay. Meanwhile, borrowers encumbered by a prepayment penalty, or borrowers using a loan to buy a condo verse a single-family detached house tend to be less likely to prepay, *ceteris paribus*.

The estimation results on the individual-level race (ethnicity) variables indicate that both African American borrowers and Hispanic borrowers are less likely to prepay than non-Hispanic White borrowers after other covariates are controlled for. These results are consistent in Model (2) in which the neighborhood-level race (ethnicity) variables are accounted for, though either of the two neighborhood-level race (ethnicity) variables is significant. One possible explanation for this finding is African American borrowers and Hispanic borrowers might face more obstacles gaining access to credit in the refinance market relative to White borrowers, though we lack data to explore this issue.

The estimation results on the contract rate determination model based on the whole pooled sample are reported in Table 3.5. Two specifications are applied. In Model (1), the three predicted loan termination probabilities are calculated based on the specifications of Model (1) in Table 3.3 and 4. In other words, neighborhood-level race (ethnicity) variables are not included to predict loan default probability and prepayment probabilities. By contrast, for Model (2) in Table 3.5, neighborhood-level race (ethnicity) variables are included to calculate the predicted loan termination probabilities. Accordingly, in Table 3.5, neighborhood-level race (ethnicity) variables are excluded in Model (1), whereas they are incorporated in Model (2). The estimation results are quite consistent in these two specifications.

The estimation results in Table 3.5 indicate that when a lender sets a loan's contract rate, the lender takes into account how likely a borrower is to default or prepay in the future; as the

coefficients on the predicted default probability, pecuniary prepayment probability, and non-pecuniary prepayment probability are all significant. This finding also implies that it is important and necessary to include these variables in the loan contract rate determination equation. The coefficient estimate on each of these variables is positive, indicating lenders charge positive risk premiums for both default risk and prepayment risk. Specifically, the increase in the contract rate from a 10-percentage-point increase in the default probability is approximately 12 basis points. A 10-percentage-point increase in the pecuniary prepayment probability would lead to an increase in the contract rate by 25 basis points, whereas the same amount of increase in the probability of non-pecuniary prepayment would result in an increase in the contract rate by 24 basis points.

The estimation results on other covariates in the loan contract rate determination equation are consistent with either theoretical expectations or prior studies. The estimate on the 10-year yield at origination is significantly positive, and close to 1. The estimates on the three category variables of original LTV ratio are all significant and positive. The results show as the original LTV ratio increases, loan contract rate rises monotonically. Meanwhile, the borrower's FICO score plays an important role in determining a loan's contract rate. Lenders are shown to charge a higher contract rate for borrowers with a lower FICO score. However, when a borrower's FICO score is above 700, we fail to reject that an additional increase in FICO score does not impact loan contract rate. In addition, lenders tend to charge a lower contract rate to borrowers who provide full income documentation at origination, as well as borrowers who used a loan to purchase a house as the primary residence rather than as investment. The type of the underlying property is shown to influence the contract rate also. Borrowers using a loan to purchase a condo appears to be charged at a higher contract rate on average than borrowers using a loan to purchase a single-family

detached house. Additionally, among the neighborhood traits variables, only the recent house price appreciation rate is shown to affect loan contract rate, in a negatively.

The estimation results on the original loan size variables and two prepayment penalty category variables require additional explanations. The results on the original loan size indicate that as the original loan size rises, the contract rate decreases, but at a decreasing rate. This finding is consistent with most of the previous related studies. One possible explanation for this finding is the original loan size is believed to be positively associated with a borrower's income and wealth, and the latter is anticipated to be negatively correlated with loan contract rate. Another possible explanation takes the loan-serving fees into account. Loan-serving fees typically do not depend on the original loan size, therefore, the loan-serving fees per unit of original loan amount decreases as the original loan amount increases. Because the loan-serving fees have been incorporated in the contract rate, a negative relationship between contract rate and original loan size is anticipated. The estimates on the two category prepayment penalty variables are shown to be significant and positive. This finding is somewhat puzzling given that prepayment penalty transfers a portion of the prepayment risk from the lender to the borrower, and thus should reduce the contract rate. However, a close examination reveals that, as the length of penalty period increases, the contract rate decreases, indicating that the penalty itself reduces loan contract rate. One possible explanation for the observed positive relationship between prepayment penalty and loan contract rate is borrowers selecting a prepayment penalty tend to be riskier, other things equal.⁸⁵

Turning our attention to the estimates on the individual- and neighborhood-level race (ethnicity) variables, an African American borrower is shown to pay a contract rate more than 30 basis points higher than a non-Hispanic White borrower, *ceteris paribus*; while a Hispanic

⁸⁵ A prior study by Mayer et al. (2013) provided evidence for this explanation.

borrower appears to pay a contract rate approximately 10 basis points higher than a White counterpart. Meanwhile we fail to reject that the racial (ethnic) composition of a borrower's neighborhood does not have any impact on a loan's contract rate after the borrower's race (ethnicity) as well as a series of covariates are accounted for as shown in Model (2). As the predicted loan termination probabilities are all controlled for, the racial (ethnic) disparity in loan contract rate demonstrated here could only be attributed to the borrower's race (ethnicity) itself.

3.6.2. Matching Results

3.6.2.1. Matching Between African American Borrowers and non-Hispanic White borrowers

Three loans by African American borrowers in the sample are not included in matching simply because they are not within the region of common support in terms of the three predicted loan termination probabilities ($\hat{P}_k$).⁸⁶ Following prior studies analyzing the optimal caliper radius ratio (ρ) for matching, this study allows the caliper radius ratio (ρ) to range from 0.2 to 1.0 in increments of 0.2.⁸⁷ An appropriate caliper radius ratio is adopted if the caliper radius ratio could significantly reduce the mean difference between matched pairs without losing too much power by excluding too many “target” loans without a successful match. For the matching between African American borrowers and non-Hispanic White borrowers, a caliper radius ratio of 1.0 and a caliper radius ratio of 0.8 turn out to perform well.⁸⁸

With a caliper radius ratio of 0.8, 22 out of 97 loans by African Americans failed to find a successful match, leaving 75 matched pairs in matching sample 1. A caliper radius ratio of 1.0

⁸⁶ The three predicted loan termination probabilities used for matching are generated based on the specifications of Model (1) in Table 3.3 and 3.4. Neighborhood-level race (ethnicity) variables are not included for the prediction, as the estimates on the two variables are all insignificant in both the default hazard model and prepayment hazard model.

⁸⁷ Cochran and Rubin (1973), Rosenbaum and Rubin (1985), and Austin (2011) studied the optimal caliper radius ratio.

⁸⁸ A caliper radius ratio less than 0.8 excluded more than half of the loans by African American borrowers.

only excluded 7 “target” loans, leaving 89 matched pairs in matching sample 2. Table B.2 in the Appendices reports the matching balance diagnosis results on matching sample 1, whereas Table B.3 in the Appendices demonstrates the matching balance diagnosis results on matching sample 2. Both matching sample 1 and matching sample 2 are well balanced with regard to the predicted probability of each event ($\hat{P}_k$), the 10-year treasury yield at origination (y_0), as well as most of those variables describing the characteristics of the loan, the borrowers, and the collateral, as the standardized difference of the mean of each variable falls in the range of (-0.25, 0.25), the ratio of the variance of each variable is within the range of (0.5, 2), and mean difference t test is not statistically significant. However, as shown in Table B.2 and Table B.3, in either matching sample 1 or matching sample 2, a few variables are not well balanced, especially those variables describing the traits of a borrower’s neighborhood.⁸⁹ Because of the few unbalanced variables in the two matching samples, a regression adjustment is applied to each matching sample to account for the differences in those unbalanced variables. Specifically, for each matching sample, the loan contract rate is regressed against all the covariates as specified in Equation (7) to test if a borrower’s race has an additional effect on loan contract rate after the predicted probability of each event ($\hat{P}_k$), the 10-year yield at origination (y_0), and all of the covariates in $\mathbf{z}$ are completely controlled for.

Table 3.6 reports the results on nearest 1-to-1 matching with replacements and with calipers between loans by African American borrowers and loans by non-Hispanic White borrowers, based on matching sample 1 and 2, respectively. In Panel A, the mean difference in loan contract rate between African American borrowers and non-Hispanic White borrowers is reported, and a two-

⁸⁹ Though most of the variables describing the traits of a borrower’s neighborhood are unbalanced, we still think each matching sample is overall balanced, because the estimates on those neighborhood variables are shown not to significantly affect loan default probability or prepayment probability.

sample t test is conducted to test if the mean difference is zero. In Panel B, the regression adjustment results are demonstrated for matching sample 1 and 2, respectively. Only the estimate on the individual-level race variable is reported, with the estimates on other covariates compressed.⁹⁰ The results show that without regression adjustment, the average contract rate charged to African American borrowers tends to be significantly higher than that by non-Hispanic White borrowers, specifically, approximately 50 basis points higher. With regression adjustment, African American borrowers are shown to pay a contract rate roughly 30 basis points higher than their non-Hispanic White counterparts on average. The results are quite consistent in these two matching samples, and the magnitude of this racial difference in loan contract rate with regression adjustment (30 basis points) is quite close to the one only with regression analysis based on the whole pooled sample as reported in Table 3.5.

3.6.2.2. Matching between Hispanic Borrowers and non-Hispanic White borrowers

All of the loans in the sample by non-Hispanic White borrowers are included for matching because all of them are within the region of the common support in terms of the three predicted loan termination probabilities ($\hat{P}_k$).⁹¹ Here, a caliper radius ratio of 0.6, 0.8, and 1.0, respectively, results in a balanced matching sample with sufficient matched pairs for analysis.⁹²

With a caliper radius ratio of 0.6, 68 out of 309 loans by non-Hispanic White borrowers could not find a successful match, leaving 241 matched pairs in matching sample 1. A caliper radius ratio of 0.8 results in a matching sample named matching sample 2 consisting of 283

⁹⁰ The estimates on other variables are omitted here to save space, but are available upon request.

⁹¹ Similar to the matching between African American borrowers and non-Hispanic White borrowers, the three predicted loan termination probabilities used for matching here are generated based on the specifications of Model (1) in Table 3.3 and 3.4. Neighborhood-level race (ethnicity) variables are not included for the prediction.

⁹² A caliper radius ratio less than 0.6 dropped more than half of the loans by non-Hispanic White borrowers, and could not make any improvements in reducing the bias between the two groups.

matched pairs. There are 298 matched pairs in matching sample 3 with a caliper radius ratio of 1.0. Matching balance diagnosis results on matching sample 1, sample 2, and sample 3, are reported in Table B.4, Table B.5, and Table B.6 in the Appendices, respectively. Only few variables are not well balanced in the three matching samples, but none of them depicts the characteristics of the loan or the borrower, and none of them is believed to be a major loan risk factor.⁹³ A regression adjustment is conducted for each of the three matching samples in the same way as we did for the matching between African American borrowers and non-Hispanic White borrowers.

Results on nearest 1-to-1 matching with replacements and with calipers between loans by Hispanic borrowers and loans by non-Hispanic White borrowers on the basis of matching sample 1, sample 2, and sample 3, respectively, are demonstrated in Table 3.7. Similar to Table 3.6, in Panel A, the mean difference in loan contract rate between the two racial (ethnic) groups is reported, and whether the mean difference is zero is tested. In Panel B of Table 3.7, the regression results of the contract rate determination equation on each of the three matching samples are reported. The results in Panel A indicate that without regression adjustment, Hispanic borrowers appear to pay a significantly higher contract rate on average than non-Hispanic White borrowers. The magnitude of this rate disparity ranges from 23 basis points to 28 basis points. In addition, as demonstrated in Panel B of Table 3.7, regression adjustment does not seem to change the results. The estimate on the individual-level race (ethnicity) variable is significantly positive in all of the three matching samples, and the magnitude of this coefficient estimate ranges from 23 basis points to 30 basis points. Recall that, the magnitude of the estimated racial (ethnic) difference in contract rate between Hispanic borrowers and non-Hispanic White borrowers based on the whole pooled

⁹³ Only one variable describes the traits of the collateral, and three variables depict the characteristics of the neighborhood. Two of the three neighborhood variables fail to have significant estimates in loan hazard model as shown in Table 3.3 and 3.4.

sample is only approximately 10 basis points as reported in Table 3.5. It is nearly half of the magnitude of the estimated racial (ethnic) disparity in contract rate based on the matching samples here. This finding confirms our concerns on the potential problems involved in regression analysis on the whole pooled sample.

Overall both the parametric results and matching results provide empirical evidence of racial (ethnic) inequality in mortgage lending. African American borrowers and Hispanic borrowers are shown to pay higher contract rates than comparable non-Hispanic White borrowers just because they are minorities.

3.7. Discussion

The empirical findings in this study undoubtedly demonstrate that lenders take a borrower's race (ethnicity) into account in determining a loan's contract rate beyond the extent to which it affects loan default probability and prepayment probability. In this study, several explanations are proposed to explain this lenders' loan pricing behavior.

One limitation of the data in this study is the number of points paid by a borrower is not observed. As points are usually used by borrowers to buy down loan contract rate, one might argue that a possible explanation for the higher contract rates paid by African American borrowers and Hispanic borrowers is they tend to choose less points on average at loan closing. However, this hypothesis seems not plausible to fully explain the racial (ethnic) disparity in loan contract rate. One interesting and prominent finding in this study is both African American borrowers and Hispanic borrowers are shown to be less likely to prepay, other things equal. Rational borrowers, anticipating a lower likelihood of prepaying a loan in the near future, would choose to pay more points. The finding in this study does not support this hypothesis, though not directly. Additionally,

prior studies (Woodward, 2008) provided empirical evidence that minority borrowers pay more in closing costs at loan origination than White counterparts. Given the fact that there is no convincing evidence showing African American borrowers and Hispanic borrowers pay less points up front, this explanation is unlikely to be the major reason for the racial (ethnic) difference in loan contract rate.

A second explanation, which may be ironic, could be the effects of regulation. According to the Equal Credit Opportunity Act (ECOA), it is illegal for lenders to consider a borrower's race (ethnicity) in their decisions on loan pricing. On the basis of the option-based pricing model of fixed-rate mortgage loans (Kau et al., 1995), both the value of the default option and the value of the prepayment option have been incorporated in the value of a mortgage loan, and following the "law of one price", an equilibrium mortgage contract rate at origination is determined as the one that yields a normal rate of return on the loan. Hence, in a competitive market without arbitrage opportunities, loans with less prepayment option values should be priced at lower contract rates, *ceteris paribus*. In this study, results on loan termination patterns indicate that African American borrowers and Hispanic borrowers tend to be less likely to exercise the prepayment option than non-Hispanic White borrowers, after other prepayment risk factors are controlled for. This indicates that, keeping other things equal, the contract rate of loans by African American borrowers or Hispanic borrowers should be lower than that by non-Hispanic White borrowers. However, the lenders, with the fear of legal or social sanctions, would ignore the truth and disregard the known effects of race (ethnicity) on loan termination patterns, especially on loan prepayment patterns. Therefore, because of the regulation, the actual contract rates charged to African American borrowers and Hispanic borrowers are higher than what should have been rationally determined in

a competitive market. The regulation ironically contributes to the racial (ethnic) disparity in loan contract rate, if not entirely, but at least partially.

Finally, the contract rate disparity across racial (ethnic) groups could simply be attributed to discrimination. It does not definitely mean that lenders practice pricing discrimination based on a borrower's race (ethnicity) consciously. It might be possible that a mortgage loan is priced by a lender under a false impression that a loan by an African American borrower or a Hispanic borrower was riskier than what it actually is. Whatever the reason is, lenders discriminate against minority borrowers.

In order to test the extent to which regulation and discrimination contribute to the racial (ethnic) disparity in loan pricing, respectively, a race-blind model is applied. In a race-blind world, it is assumed that lenders deliberately pretend that they do not know the role of race (ethnicity) in loan prepayment patterns. Therefore, in this model, loan termination behavior model estimation remains the same with the individual-level race fixed effects variables included, however, when calculating the predicted prepayment probabilities, African American borrowers and Hispanic borrowers are treated as they were non-Hispanic White borrowers by replacing the individual-level race fixed effects variables with zeros. In this way, individual-level race (ethnicity) does not contribute to the predicted prepayment probabilities, and the "race-blind" predicted prepayment probabilities are actually higher than the true predicted prepayment probabilities for both African American borrowers and Hispanic borrowers. The resulting "race-blind" predicted prepayment probabilities as well as the predicted default probability are included in the loan contract rate determination equation to investigate if a borrower's race (ethnicity) has a separate role on a loan's contract rate in a "race-blind" world. As the "race-blind" predicted prepayment probabilities have incorporated the contribution of regulation, if there is a separate role played by a borrower's race

(ethnicity) in the contract rate equation, that separate role is entirely because of discrimination, and the remaining part is due to regulation.⁹⁴ Table 3.8 reports the results of this “race-blind” model based on the pooled sample. The results show that, one individual-level race fixed effects variable – *African Americans* – still remains significant but the coefficient estimate appears to be less than the one in a competitive world as shown in Table 3.5, whereas the other race fixed effects variable – *Hispanics* – becomes insignificant. Those results indicate that the higher contract rates paid by Hispanic borrowers as shown in Table 3.5 are entirely attributable to regulation, while both regulation and discrimination explain the higher contract rates paid by African American borrowers. Specifically, in the pooled sample, African American borrowers on average tend to pay a contract rate approximately 30 basis points higher than White counterparts as shown in Table 3.5, out of the additional 30 basis points, 18 basis points are attributable to discrimination, while the remaining 12 basis points are due to regulation.

3.8. Conclusion

This study investigates the impact of race (ethnicity) on the pricing of subprime mortgage loans based on a sample of 30-year fixed-rate home purchase loans originated from 1997 to 2006 in Miami-Dade County, FL. Given the consideration that race (ethnicity) may be correlated with loan termination behavior, this study explicitly models and controls for loan default probability and prepayment probability. Within a parametric analysis framework, the predicted loan default

⁹⁴ Another approach to separate the contribution of regulation and discrimination is to compare the “race-blind” predicted prepayment probabilities and the true predicted prepayment probabilities for each loan by an African American borrower or a Hispanic borrower in the pooled sample, and to test whether the mean difference is significantly different from zero for both pecuniary prepayment and non-pecuniary prepayment. The mean difference in each race (ethnicity) group between the “race-blind” predicted pecuniary/non-pecuniary prepayment probability and the true one is further multiplied by the estimated pecuniary/non-pecuniary prepayment risk premium in Table 3.5 (the coefficient estimate on pecuniary/non-pecuniary prepayment probability). The resulting two products are added together in order to calculate the extent to which regulation contributes to the racial disparity in contract rates. The results by this approach are quite close to the results reported in Table 3.8, and are omitted here.

probability and prepayment probability are included in a loan contract rate determination model along with individual- and neighborhood-level race (ethnicity) variables. Under a matching analysis framework, the predicted default probability and prepayment probability are used to match loans in two racial (ethnic) groups to remove any systematic differences in loan termination risk between the two groups.

The results provide empirical evidence of racial inequality in the subprime mortgage market over the study period. Results on loan hazard assessment show a borrower's race (ethnicity) is associated with loan prepayment probability. Both African American borrowers and Hispanic borrowers are shown to be less likely to prepay than non-Hispanic White borrowers. When examining the relationship between loan contract rate and race, both the regression results based on the pooled sample and the matching results based on each matching sample provide evidence of adverse pricing against African American borrowers and Hispanic borrowers. The results indicate that the average contract rates paid by African American borrowers is approximately 30 basis points higher than that by non-Hispanic White counterparts, and the rate disparity between Hispanic borrowers and non-Hispanic White borrowers ranges from 23 basis points to 30 basis points after matching. The significant higher contract rates paid by African American borrowers and Hispanic borrowers could only be attributed to race (ethnicity) itself.

Several possible explanations are proposed to explain the observed contract rate disparity across racial (ethnic) groups. It is ironic that both discrimination and regulation contribute to the observed adverse pricing against minorities. The results indicate that the observed rate disparity between Hispanic borrowers and non-Hispanic White borrowers is entirely attributable to regulation, while nearly half of the observed rate disparity between African American borrowers

and non-Hispanic White borrowers is explained by regulation, with the remaining part due to discrimination.

References

- Allison, Paul D. *Survival analysis using SAS: a practical guide*. SAS Institute, 2010.
- An, Xudong, John M. Clapp, and Yongheng Deng. "Omitted mobility characteristics and property market dynamics: application to mortgage termination." *The Journal of Real Estate Finance and Economics* 41, no. 3 (2010): 245-271.
- Austin, Peter C. "Optimal caliper widths for propensity-score matching when estimating differences in means and differences in proportions in observational studies." *Pharmaceutical statistics* 10, no. 2 (2011): 150-161.
- Avery, Robert, Kenneth Brevoort, and Glenn Canner. "Opportunities and issues in using HMDA data." *Journal of Real Estate Research* (2009).
- Black, Dan A., Amelia M. Haviland, Seth G. Sanders, and Lowell J. Taylor. "Gender wage disparities among the highly educated." *Journal of human resources* 43, no. 3 (2008): 630-659.
- Black, Harold A., Thomas P. Boehm, and Ramon P. DeGennaro. "Is there discrimination in mortgage pricing? The case of overages." *Journal of Banking & Finance* 27, no. 6 (2003): 1139-1165.
- Boehm, Thomas P., Paul D. Thistle, and Alan Schlottmann. "Rates and race: An analysis of racial disparities in mortgage rates." *Housing Policy Debate* 17, no. 1 (2006): 109-149.
- Cheng, Ping, Zhenguo Lin, and Yingchun Liu. "Racial discrepancy in mortgage interest rates." *The Journal of Real Estate Finance and Economics* 51, no. 1 (2015): 101-120.
- Clapp, John M., Gerson M. Goldberg, John P. Harding, and Michael LaCour-Little. "Movers and shuckers: interdependent prepayment decisions." *Real estate economics* 29, no. 3 (2001): 411-450.
- Cochran, William G., and Donald B. Rubin. "Controlling bias in observational studies: A review." *Sankhyā: The Indian Journal of Statistics, Series A* (1973): 417-446.
- Crawford, Gordon W., and Eric Rosenblatt. "Differences in the cost of mortgage credit implications for discrimination." *The Journal of Real Estate Finance and Economics* 19, no. 2 (1999): 147-159.
- Courchane, Marsha. "The Pricing of Home Mortgage Loans to Minority Borrowers: How Much of the APR Differential Can We Explain?" *Journal of Real Estate Research* (2009).
- Cox, John C., Jonathan E. Ingersoll Jr, and Stephen A. Ross. "A theory of the term structure of interest rates." *Econometrica: Journal of the Econometric Society* (1985): 385-407.
- Deng, Yongheng, John M. Quigley, and Robert Order. "Mortgage terminations, heterogeneity and the exercise of mortgage options." *Econometrica* 68, no. 2 (2000): 275-307.

- Deng, Yongheng, and Stuart A. Gabriel. "Risk-based pricing and the enhancement of mortgage credit availability among underserved and higher credit-risk populations." *Journal of Money, Credit, and Banking* 38, no. 6 (2006): 1431-1460.
- Duan, Jin-Chuan, and Jean-Guy Simonato. "Estimating and testing exponential-affine term structure models by Kalman filter." *Review of Quantitative Finance and Accounting* 13, no. 2 (1999): 111-135.
- Duca, John V., and Stuart S. Rosenthal. "Do mortgage rates vary based on household default characteristics? Evidence on rate sorting and credit rationing." *The Journal of Real Estate Finance and Economics* 8, no. 2 (1994): 99-113.
- Firestone, Simon, Robert Van Order, and Peter Zorn. "The Performance of Low-Income and Minority Mortgages." *Real Estate Economics* 35, no. 4 (2007): 479-504.
- Frölich, Markus. "Propensity score matching without conditional independence assumption—with an application to the gender wage gap in the United Kingdom." *The Econometrics Journal* 10, no. 2 (2007): 359-407.
- Getter, Darryl E. "Consumer credit risk and pricing." *Journal of Consumer Affairs* 40, no. 1 (2006): 41-63.
- Ghent, Andra C., Rubén Hernández-Murillo, and Michael T. Owyang. "Differences in subprime loan pricing across races and neighborhoods." *Regional Science and Urban Economics* 48 (2014): 199-215.
- Kau, James B., Donald C. Keenan, and Henry J. Munneke. "Racial discrimination and mortgage lending." *The Journal of Real Estate Finance and Economics* 45, no. 2 (2012): 289-304.
- Kau, James B., Donald C. Keenan, Walter J. Muller III, and James F. Epperson. "The valuation at origination of fixed-rate mortgages with default and prepayment." *The Journal of Real Estate Finance and Economics* 11, no. 1 (1995): 5-36.
- Kelly, Austin. "Racial and ethnic disparities in mortgage prepayment." *Journal of Housing Economics* 4, no. 4 (1995): 350-372.
- LaCour-Little, Michael. "Discrimination in mortgage lending: A critical review of the literature." *Journal of Real Estate Literature* 7, no. 1 (1999): 15-50.
- Ladd, Helen F. "Evidence on discrimination in mortgage lending." *The Journal of Economic Perspectives* 12, no. 2 (1998): 41-62.
- Ñopo, Hugo. "Matching as a tool to decompose wage gaps." *The review of economics and statistics* 90, no. 2 (2008): 290-299.
- Nothaft, Frank E., and Vanessa G. Perry. "Do mortgage rates vary by neighborhood? Implications for loan pricing and redlining." *Journal of Housing Economics* 11, no. 3 (2002): 244-265.

- Mayer, Chris, Tomasz Piskorski, and Alexei Tchistyi. "The inefficiency of refinancing: Why prepayment penalties are good for risky borrowers." *Journal of Financial Economics* 107, no. 3 (2013): 694-714.
- Rosenbaum, Paul R., and Donald B. Rubin. "The central role of the propensity score in observational studies for causal effects." *Biometrika* 70, no. 1 (1983): 41-55.
- Rosenbaum, Paul R., and Donald B. Rubin. "Constructing a control group using multivariate matched sampling methods that incorporate the propensity score." *The American Statistician* 39, no. 1 (1985): 33-38.
- Ross, Stephen L., and John Yinger. "The color of credit: Mortgage discrimination, research methodology, and fair-lending enforcement." *MIT Press Books* 1 (2002).
- Rubin, Donald B., and Neal Thomas. "Combining propensity score matching with additional adjustments for prognostic covariates." *Journal of the American Statistical Association* 95, no. 450 (2000): 573-585.
- Turner, Margery Austin and Skidmore, Felicity (Editors). 1999. "Mortgage Lending Discrimination: A Review of Existing Evidence." The Urban Institute.
- Woodward, SusanE. "A Study of Closing Costs for FHA Mortgages. US Department of Housing and Urban Development." *Office of Policy Development and Research* (2008).
- Zhang, Yan. "Fair lending analysis of mortgage pricing: does underwriting matter?" *The Journal of Real Estate Finance and Economics* 46, no. 1 (2013): 131-151.

Table 3.1 Descriptive Statistics of Mortgage Loans at Loan Origination

Variable Name	Mean	Std. Dev.	Min.	Max.
Default (0,1)	0.0869	0.2818	0.0000	1.0000
Prepay (0,1)	0.8041	0.3970	0.0000	1.0000
Loan Characteristics				
Contract rate at origination (C_0)	7.9622	1.1282	5.2500	12.5000
10-year treasury yield at origination (y_0)	4.9201	0.8163	3.3300	6.8900
Contract rate spread at origination ($C_0 - y_0$)	3.0422	0.9476	1.0100	6.4650
Original LTV	84.8048	11.8555	36.0000	107.0000
FICO at origination	700.0370	53.2957	483.0000	822.0000
Original loan amount (in \$10,000)	18.1243	12.1792	2.0000	80.0000
Full income documentation (0,1)	0.4765	0.4996	0.0000	1.0000
Without prepayment penalty (0,1)	0.6702	0.4703	0.0000	1.0000
Prepayment penalty for 1 to 3 years (0,1)	0.2236	0.4168	0.0000	1.0000
Prepayment penalty for 5 years (0,1)	0.1061	0.3081	0.0000	1.0000
Property Characteristics				
Property owner occupied (0,1)	0.7778	0.4159	0.0000	1.0000
Property condo (0,1)	0.3376	0.4731	0.0000	1.0000
Neighborhood-Level Characteristics at Origination ^a				
Recent housing price appreciation rate ^b	0.1949	0.2227	-0.4519	2.8150
Heterogeneity in housing price (in \$10,000) ^c	11.2656	7.8752	1.1012	43.3243
Housing occupancy rate (from Census Survey) ^d	0.9094	0.0858	0.5373	0.9883
Poverty rate (from Census Survey) ^d	0.1441	0.0826	0.0216	0.5988
Median applicant income (from HMDA) ^e	1.2188	0.5786	0.5483	4.8597
Proportion of non-Hispanic White applicants (from HMDA) ^f	0.2069	0.1729	0.0171	0.7069
Proportion of Hispanic applicants (from HMDA) ^f	0.6902	0.2161	0.0742	0.9778
Proportion of African American applicants (from HMDA) ^f	0.0775	0.1439	0.0000	0.8399
Sample Size	1,404			

^a The neighborhood of the property as the collateral of a loan is defined as the property's location based on the 1990 census tract boundaries.

^b Recent housing price appreciation rate at origination is defined as the ratio of the median housing sale price in a census tract in a three-year period prior to the month of loan origination to the median housing sale price in the same census tract in another three-year period prior to the three-year pre-origination period, then minus 1.

^c Heterogeneity in housing price at origination is defined as the standard deviation of the housing sale price in a census tract over a three-year period prior to the month of loan origination.

^d Housing occupancy rate and poverty rate were generated from the decennial census survey data in 1990, 2000, and 2010.

^e Median applicant income was generated from the HMDA data aggregated on the census tract level on a yearly basis. It is defined as the ratio of the median applicant income in a census tract at origination to the median applicant income in Miami-Dade County, FL at origination.

^f Variables on the racial and ethnic composition of a census tract were generated from the HMDA data aggregated on the census tract level on a yearly basis.

Table 3.2 Descriptive Statistics of Mortgage Loans at Loan Origination by Race

Variable Name	Race Group		
	Non-Hispanic White	Hispanic	African American
	Mean	Mean	Mean
Default (0,1)	0.0647	0.0894	0.1300
Prepay (0,1)	0.8803	0.7839	0.7700
Loan Characteristics			
Contract rate at origination (C_0)	7.9198	7.9140	8.5731
10-year treasury yield at origination (y_0)	5.1425	4.8526	4.9040
Contract rate spread at origination ($C_0 - y_0$)	2.7773	3.0614	3.6691
Original LTV	81.2395	85.4995	88.9100
FICO at origination	712.7605	698.0683	680.3100
Original loan amount (in \$10,000)	22.0816	17.4520	12.5855
Full income documentation (0,1)	0.4822	0.4633	0.5900
Without prepayment penalty (0,1)	0.2201	0.3497	0.4700
Prepayment penalty for 1 to 3 years (0,1)	0.0939	0.2492	0.3700
Prepayment penalty for 5 years (0,1)	0.1262	0.1005	0.1000
Property Characteristics			
Property owner occupied (0,1)	0.7767	0.7799	0.7600
Property condo (0,1)	0.3495	0.3508	0.1700
Neighborhood-Level Characteristics at Origination ^a			
Recent housing price appreciation rate ^b	0.1815	0.2016	0.1693
Heterogeneity in housing price (in \$10,000) ^c	15.8298	10.2828	6.9411
Housing occupancy rate (from Census Survey) ^d	0.8748	0.9196	0.9148
Poverty rate (from Census Survey) ^d	0.1314	0.1410	0.2142
Median applicant income (from HMDA) ^e	1.5494	1.1506	0.8755
Proportion of non-Hispanic White applicants (from HMDA) ^f	0.3545	0.1676	0.1415
Proportion of Hispanic applicants (from HMDA) ^f	0.5469	0.7567	0.4708
Proportion of African American applicants (from HMDA) ^f	0.0646	0.0533	0.3587
Number of loans	309	995	100

^a The neighborhood of the property as the collateral of a loan is defined as the property's location based on the 1990 census tract boundaries.

^b Recent housing price appreciation rate at origination is defined as the ratio of the median housing sale price in a census tract in a three-year period prior to the month of loan origination to the median housing sale price in the same census tract in another three-year period prior to the three-year pre-origination period, then minus 1.

^c Heterogeneity in housing price at origination is defined as the standard deviation of the housing sale price in a census tract over a three-year period prior to the month of loan origination.

^d Housing occupancy rate and poverty rate were generated from the decennial census survey data in 1990, 2000, and 2010.

^e Median applicant income was generated from the HMDA data aggregated on the census tract level on a yearly basis. It is defined as the ratio of the median applicant income in a census tract at origination to the median applicant income in Miami-Dade County, FL at origination.

^f Variables on the racial and ethnic composition of a census tract were generated from the HMDA data aggregated on the census tract level on a yearly basis.

Table 3.3 Default Hazard Model Estimates

Variables	Model (1)		Model (2)	
	Coef.	P-value	Coef.	P-value
Intercept	6.7848	0.1509	6.4472	0.1871
Contract rate spread at origination ($C_0 - y_0$)	0.0382	0.9095	0.0565	0.8660
Market interest rate change at time t ($y_0 - y_{t-2}$)	-0.3202	0.5411	-0.3274	0.5332
Relative house price at time t (RHP_t) ^a	-2.9313	<0.0001	-2.9636	<0.0001
$(y_0 - y_{t-2}) \times (RHP_t)$ ^a	0.6058	0.2051	0.6122	0.2020
Original LTV categories (Base group: Original LTV ≤ 80)				
80 < Original LTV ≤ 90	0.0273	0.9241	0.0231	0.9356
90 < Original LTV ≤ 100	0.5549	0.1491	0.5379	0.1627
100 < Original LTV	1.2497	0.0373	1.2272	0.0423
FICO at origination continuous linear splines:				
Minimum (FICO, 700)	-0.0130	0.0088	-0.0126	0.0114
Maximum (FICO, 700) – 700	-0.0004	0.9147	-0.0005	0.8960
Original loan amount (in \$10,000)	0.1352	<0.0001	0.1378	<0.0001
Square term of original loan amount (in \$10,000)	-0.0012	0.0131	-0.0012	0.0128
Full income documentation (0,1)	-0.6783	0.0047	-0.6791	0.0049
Property owner occupied (0,1)	-0.5630	0.0257	-0.5632	0.0265
Property condo (0,1)	0.3131	0.2245	0.3315	0.2197
Neighborhood-level heterogeneity in housing price at time t (in \$10,000) ^a	0.0301	0.1891	0.0285	0.2367
Neighborhood-level housing occupancy rate at time t ^a	-2.1707	0.1284	-1.9855	0.1832
Neighborhood-level poverty rate at time t ^a	1.6023	0.2588	1.5569	0.2772
Neighborhood-level median applicant income at time t (HMDA) ^a	-1.7403	0.0001	-1.7441	0.0002
Neighborhood-level proportion of African Americans at time t (HMDA) ^a			-0.0234	0.9864
Neighborhood-level proportion of Hispanics at time t (HMDA) ^a			-0.2288	0.8156
Borrower Race (Base group: Non-Hispanic Whites)				
African Americans	0.2983	0.4495	0.2831	0.5033
Hispanics	-0.0768	0.7714	-0.0603	0.8228
Residual ^b	0.6739	0.0549	0.6561	0.0611
SDA ^a	1.0343	0.0045	1.0497	0.0043
Origination season fixed effects ^c	YES		YES	
Likelihood Ratio	230.9555		230.8994	

^a Demote time-varying variables.^b The residual comes from the contract-rate reduced-form estimation.^c Loan origination season fixed effects estimates are omitted here, but are available upon request.

Table 3.4 Prepay Hazard Model Estimates

Variables	Model (1)		Model (2)	
	Coef.	P-value	Coef.	P-value
Intercept	-10.9079	<0.0001	-10.8479	<0.0001
Contract rate spread at origination ($C_0 - y_0$)	0.4954	<0.0001	0.4834	<0.0001
Market interest rate change at time t ($y_0 - y_{t-2}$) ^a	0.5781	0.0002	0.5741	0.0002
Relative house price at time t (RHP) _{t} ^a	1.3052	<0.0001	1.3042	<0.0001
($y_0 - y_{t-2}$) $\times$ (RHP) _{t} ^a	-0.0763	0.5119	-0.0734	0.5298
FICO at origination continuous linear splines:				
Minimum (FICO, 700)	0.0045	0.0016	0.0044	0.0020
Maximum (FICO, 700) – 700	-0.0021	0.1048	-0.0021	0.1001
Original loan amount (in \$10,000)	0.0298	0.0046	0.0301	0.0046
Square term of original loan amount (in \$10,000)	-0.0004	0.0117	-0.0005	0.0111
Within prepayment penalty period at time t (0,1) ^a	-0.5112	<0.0001	-0.5057	<0.0001
Property owner occupied (0,1)	0.0201	0.8089	0.0206	0.8047
Property condo (0,1)	-0.2203	0.0124	-0.2132	0.0183
Neighborhood-level housing occupancy rate at time t ^a	0.3442	0.4362	0.3369	0.4778
Neighborhood-level poverty rate at time t ^a	-0.3385	0.4861	-0.3579	0.4657
Neighborhood-level average loan applicant income at time t (HMDA) ^a	0.1344	0.0644	0.1426	0.0964
Neighborhood-level proportion of African Americans at time t (HMDA) ^a			0.1357	0.7342
Neighborhood-level proportion of Hispanics at time t (HMDA) ^a			0.0402	0.8874
Borrower Race (Base group: Non-Hispanic Whites)				
African Americans	-0.3069	0.0338	-0.3291	0.0400
Hispanics	-0.2456	0.0018	-0.2465	0.0036
Residual ^b	-0.0622	0.5602	-0.0479	0.6555
Mortgage year fixed effects ^c	YES		YES	
Origination season fixed effects ^c	YES		YES	
Likelihood Ratio	620.2292		620.2642	

^a Demote time-varying variables.

^b The residual comes from the contract-rate reduced-form estimation.

^c Loan origination season fixed effects estimates are omitted here, but are available upon request.

Table 3.5 Loan Contract Rate Estimates on the Pooled Sample -2SLS

Variables	Model (1) ^a		Model (2) ^b	
	Coef.	P-value ^c	Coef.	P-value ^c
Intercept	9.8067	<0.0001	9.7645	<0.0001
Predicted default probability	1.2160	<0.0001	1.2241	<0.0001
Predicted pecuniary prepayment probability	2.4055	<0.0001	2.4682	<0.0001
Predicted non-pecuniary prepayment probability	2.2983	<0.0001	2.3603	<0.0001
10-year treasury yield at origination (y ₀)	0.7946	<0.0001	0.7937	<0.0001
Original LTV categories (Base group: Original LTV≤80)				
80<Original LTV≤90	0.1631	<0.0001	0.1645	<0.0001
90<Original LTV≤100	0.3884	<0.0001	0.3892	<0.0001
100<Original LTV	0.5961	<0.0001	0.5990	<0.0001
FICO at origination continuous linear splines				
Minimum (FICO, 700)	-0.0104	<0.0001	-0.0104	<0.0001
Maximum (FICO, 700) – 700	0.0001	0.7873	0.0002	0.6853
Original loan amount (in \$10,000)	-0.0342	<0.0001	-0.0351	<0.0001
Square term of original loan amount (in \$10,000)	0.0005	<0.0001	0.0005	<0.0001
Full income documentation (0,1)	-0.1202	0.0016	-0.1179	0.0019
Prepayment penalty categories (Base group: No prepayment penalty)				
Prepayment penalty for 1-3 years	0.5981	<0.0001	0.6015	<0.0001
Prepayment penalty for 5 years	0.4235	<0.0001	0.4301	<0.0001
Property owner occupied (0,1)	-0.1090	0.0059	-0.1093	0.0057
Property condo (0,1)	0.1476	0.0007	0.1387	0.0019
Neighborhood-level past housing price appreciation rate at origination	-0.1710	0.0574	-0.1671	0.0646
Neighborhood-level heterogeneity in housing price at origination (in \$10,000)	-0.0022	0.5677	-0.0009	0.8309
Neighborhood-level housing occupancy rate at origination	-0.3203	0.1226	-0.4133	0.0571
Neighborhood-level poverty rate at origination	0.1113	0.6292	0.1763	0.4491
Neighborhood-level median income at origination (HMDA)	-0.0430	0.3501	-0.0363	0.4349
Neighborhood-level proportion of African Americans at origination (HMDA)			-0.0217	0.9071
Neighborhood-level proportion of Hispanics at origination (HMDA)			0.1664	0.1907
Borrower Race (Base group: Non-Hispanic Whites)				
African Americans	0.3178	<0.0001	0.3513	<0.0001
Hispanics	0.1134	0.0031	0.0956	0.0162
Time Trend	0.0025	0.8344	0.0020	0.8622
Loan origination season fixed effects ^d	YES		YES	
Number of loans	1404		1404	
Adjusted R ²	0.7645		0.7653	

^a In Model (1), the predicted probability of each event is generated by loan hazard model based on the specifications of Model (1) in Table 3.3 and 3.4.

^b In Model (2), the predicted probability of each event is generated by loan hazard model based on the specifications of Model (2) in Table 3.3 and 3.4.

^c P-values were calculated using standard-error corrections because predicted termination probabilities were generated variables. The Standard way to correct the standard error demonstrated in Appendix 6A of Wooldridge (2010) was employed.

^d Loan origination season fixed effects estimates are omitted here, but are available upon request.

Table 3.6 Matching Results on Nearest 1-to-1 Matching between African Americans and Non-Hispanic Whites ^a

Panel A Mean Difference in Loan Contract Rate without Regression Adjustment								
Variable	Matching Sample 1				Matching Sample 2			
	Matching caliper radius ratio (ρ): 0.8 ^b				Matching caliper radius ratio (ρ): 1.0 ^b			
	Mean (African American)	Mean (non- Hispanic White)	Mean Diff.	Mean Diff. t test	Mean (African American)	Mean (non- Hispanic White)	Mean Diff.	Mean Diff. t test
	Contract rate at origination	8.5799	8.0658	0.5141	2.6300	8.6031	8.0863	0.5168
Number of matched pairs		75			89			
Panel B Loan Contract Rate Estimates on Matched Samples -2SLS								
Variables	Matching Sample 1		Matching Sample 2					
	Coef.	P -value ^c	Coef.	P -value ^c				
Borrower Race (Base group: Non-Hispanic Whites)								
African Americans	0.3496	0.0006	0.3077	0.0003				
Predicted loan termination probabilities ^d	YES	YES	YES	YES				
10-year treasury yield at origination (y_0) ^d	YES	YES	YES	YES				
Loan characteristics ^d	YES	YES	YES	YES				
Borrower characteristics ^d	YES	YES	YES	YES				
Collateral characteristics ^d	YES	YES	YES	YES				
Neighborhood characteristics ^d	YES	YES	YES	YES				
Time Trend ^d	YES	YES	YES	YES				
Number of loans	150		178					
Adjusted R ²	0.7824		0.8034					

^a Nearest 1-to-1 matching with replacements and with calipers is used to match loans of an African American borrower to loans of a non-Hispanic White borrower. For each loan of an African American borrower, a loan of a non-Hispanic White borrower in the pool with the shortest Mahalanobis distance on the three predicted loan termination probabilities is selected as the sole and best match.

^b A caliper is applied to impose a tolerance level for the maximum distance on a matching criterion variable between borrowers in the two racial (ethnic) groups. Those matching criterion variables include the predicted probability of each event, the 10-year yield at origination, as well as the original LTV ratio, FICO score, and original loan amount. The caliper radius (the width of the caliper) on a matching criterion variable is specified as a ratio (ρ) of the pooled standard deviation of that variable.

^c *P-values* were calculated using standard-error corrections because predicted termination probabilities were generated variables. The Standard way to correct the standard error demonstrated in Appendix 6A of Wooldridge (2010) was employed.

^d Estimates on all of the variables in the loan contract rate determination equation except for the individual race variable are omitted here, but are available upon request.

Table 3.7 Matching Results on Nearest 1-to-1 Matching between Hispanics and Non-Hispanic Whites ^a

Panel A Mean Difference in Loan Contract Rate without Regression Adjustment												
Variable	Matching Sample 1				Matching Sample 2				Matching Sample 3			
	Matching caliper radius ratio (ρ): 0.6 ^b				Matching caliper radius ratio (ρ): 0.8 ^b				Matching caliper radius ratio (ρ): 1.0 ^b			
	Mean		Mean		Mean		Mean		Mean		Mean	
	(Hispanic)	(non-Hispanic White)	Diff.	Diff. <i>t</i> test	(Hispanic)	(non-Hispanic White)	Diff.	Diff. <i>t</i> test	(Hispanic)	(non-Hispanic White)	Diff.	Diff. <i>t</i> test
	Mean	Mean	Mean	Mean	Mean	Mean	Mean	Mean	Mean	Mean	Mean	Mean
Contract rate at origination	8.2168	7.9840	0.2328	2.4700	8.1880	7.9248	0.2633	2.9300	8.1896	7.9126	0.2770	3.1500
Number of matched pairs	241				283				298			
Panel B Loan Contract Rate Estimates on Matched Samples -2SLS												
Variables	Matching Sample 1		Matching Sample 2		Matching Sample 3							
	Coef.	<i>P-value</i> ^c	Coef.	<i>P-value</i> ^c	Coef.	<i>P-value</i> ^c						
Borrower Race (Base group: Non-Hispanic Whites)												
Hispanics	0.2328	<0.0001	0.2815	<0.0001	0.3040	<0.0001						
Predicted loan termination probabilities ^d	YES	YES	YES	YES	YES	YES						
10-year treasury yield at origination (y_0) ^d	YES	YES	YES	YES	YES	YES						
Loan characteristics ^d	YES	YES	YES	YES	YES	YES						
Borrower characteristics ^d	YES	YES	YES	YES	YES	YES						
Collateral characteristics ^d	YES	YES	YES	YES	YES	YES						
Neighborhood characteristics ^d	YES	YES	YES	YES	YES	YES						
Time Trend ^d	YES	YES	YES	YES	YES	YES						
Number of loans	482		566		596							
Adjusted R ²	0.7734		0.7890		0.7664							

^a Nearest 1-to-1 matching with replacements and with calipers is used to match loans of a Hispanic borrower to loans of a non-Hispanic White borrower. For each loan of a Hispanic borrower, a loan of a non-Hispanic White borrower in the pool with the shortest Mahalanobis distance on the three predicted loan termination probabilities is selected as the sole and best match.

^b A caliper is applied to impose a tolerance level for the maximum distance on a matching criterion variable between borrowers in the two racial (ethnic) groups. Those matching criterion variables include the predicted probability of each event, the 10-year yield at origination, as well as the original LTV ratio, FICO score, and original loan amount. The caliper radius (the width of the caliper) on a matching criterion variable is specified as a ratio (ρ) of the pooled standard deviation of that variable.

^c *P-values* were calculated using standard-error corrections because predicted termination probabilities were generated variables. The Standard way to correct the standard error demonstrated in Appendix 6A of Wooldridge (2010) was employed.

^d Estimates on all of the variables in the loan contract rate determination equation except for the individual race variable are omitted here, but are available upon request.

Table 3.8 Loan Contract Rate Estimates on the Pooled Sample in a Race-Blind World - 2SLS

Variables	Coef.	P-value ^a
Intercept	9.6620	<0.0001
Predicted default probability	1.5645	<0.0001
Predicted pecuniary prepayment probability	2.6918	<0.0001
Predicted non-pecuniary prepayment probability	2.5079	<0.0001
10-year treasury yield at origination (y_0)	0.7920	<0.0001
Original LTV categories (Base group: Original LTV≤80)		
80<Original LTV≤90	0.1648	<0.0001
90<Original LTV≤100	0.3909	<0.0001
100<Original LTV	0.6006	<0.0001
FICO at origination continuous linear splines		
Minimum (FICO, 700)	-0.0104	<0.0001
Maximum (FICO, 700) – 700	0.0001	0.7982
Original loan amount (in \$10,000)	-0.0357	<0.0001
Square term of original loan amount (in \$10,000)	0.0005	<0.0001
Full income documentation (0,1)	-0.1177	0.0023
Prepayment penalty categories (Base group: No prepayment penalty)		
Prepayment penalty for 1-3 years	0.6075	<0.0001
Prepayment penalty for 5 years	0.4302	<0.0001
Property owner occupied (0,1)	-0.1025	0.0101
Property condo (0,1)	0.1364	0.0020
Neighborhood-level past housing price appreciation rate at origination	-0.1934	0.0367
Neighborhood-level heterogeneity in housing price at origination (in \$10,000)	-0.0024	0.5370
Neighborhood-level housing occupancy rate at origination	-0.3430	0.0993
Neighborhood-level poverty rate at origination	0.0962	0.6799
Neighborhood-level median income at origination (HMDA)	-0.0370	0.4275
Borrower Race (Base group: Non-Hispanic Whites)		
African Americans	0.1807	0.0133
Hispanics	0.0014	0.9669
Time Trend	0.0025	0.8336
Loan origination season fixed effects ^b	YES	
Number of loans	1404	
Adjusted R ²	0.7590	

^a *P-values* were calculated using standard-error corrections because predicted termination probabilities were generated variables. The Standard way to correct the standard error demonstrated in Appendix 6A of Wooldridge (2010) was employed.

^b Loan origination season fixed effects estimates are omitted here, but are available upon request.

Appendices

Table B.1 Definition of Variables

Variables	Variable Definition
Loan Characteristics	
Contract rate at origination (C_0)	Contract interest rate at origination, in percentage
10-year treasury yield at origination (y_0)	10-year treasury constant maturity yield at origination, in percentage
Contract rate spread at origination ($C_0 - y_0$)	Contract interest rate at origination minus 10-year treasury constant maturity yield at origination, in percentage
Original LTV	The LTV ratio at loan origination, in percentage
FICO at origination	Credit score at loan origination
Original loan amount (in \$10,000)	Loan size at loan origination (in \$10,000)
Full income documentation (0,1)	=1 for a loan with full documentation of a borrower's income
Prepayment penalty for 1-3 years	=1 for a loan encumbered by a prepayment penalty for up to the first 3 years
Prepayment penalty for 5 years	=1 for a loan encumbered by a prepayment penalty for the first 5 years
Within prepayment penalty period (0,1) ^a	=1 for a loan encumbered by a prepayment penalty in a given mortgage month t
Property Characteristics	
Property owner occupied (0,1)	=1 for a loan secured by a property occupied by the owner
Property condo (0,1)	=1 for a loan secured by a condo
Relative house price at time t (RHP) _{t} ^a	The ratio of the median house sale price index at time t to the median house sale price index at loan origination ^b

Table B.1 Definition of Variables (Continued)

Variables	Variable Definition
Neighborhood-Level Characteristics	
Recent housing price appreciation rate at origination	The ratio of the median housing sale price in a census tract in a three-year period prior to the month of loan origination to the median housing sale price in the same census tract in another three-year period prior to the three-year pre-origination period, then minus 1
Heterogeneity in housing price ^c	The standard deviation of the housing sale price in a census tract over a three-year period prior to a given month (in \$10,000)
Housing occupancy rate (from Census Survey) ^c	Housing occupancy rate in a census tract
Poverty rate (from Census Survey) ^c	Poverty rate in a census tract
Median applicant income ^c	Ratio of the median applicant income in a census tract to the median applicant income in Miami-Dade County, FL.
Proportion of non-Hispanic White applicants (from HMDA) ^c	Ratio of the number of loan applications of non-Hispanic White borrowers in a census tract to the total number of loan applications in that census tract.
Proportion of Hispanic applicants (from HMDA) ^c	Ratio of the number of loan applications of Hispanic borrowers in a census tract to the total number of loan applications in that census tract.
Proportion of African American applicants (from HMDA) ^c	Ratio of the number of loan applications of African American borrowers in a census tract to the total number of loan applications in that census tract.
Interest Rate Environment	
Market interest rate change at time t ($y_0 - y_{t-2}$) ^a	10-year treasury yield at loan origination minus 10-year treasury yield at time t , lagged by 2 months, in percentage
Others	
Time Trend	A variable that equals to 1 for a loan originated in year 1997, 2 for a loan originated in year 1998, and so forth

^a Denotes time-varying variables in default/prepayment hazard model.

^b This median house price index is at the census tract level, and is generated from the property transaction data base in Miami-Dade County, FL. For each month, this index is calculated based on the inflation-adjusted median housing sale price over a three-year window around that month, eighteen months before and eighteen months after that given month. All prices are defined in 2009 dollars.

^c Denotes variables that are time-varying variables in default/prepayment hazard model, but represent values at loan origination in the contract rate determination model.

Table B.2 Matching Balance Diagnosis on Matching Sample 1 between African Americans and Non-Hispanic Whites ^a

Variables	Mean		Std. Dev.		Mean diff.	Std. mean diff. ^b	Ratio of the var. ^b	Mean diff. <i>t</i> -test ^b
	African American	Non-Hispanic White	African American	Non-Hispanic White				
10-year treasury yield at origination (y_0)	4.9773	4.9628	0.7747	0.8090	0.0145	0.0183	0.9171	0.1100
Predicted default probability	0.0982	0.0941	0.0901	0.0969	0.0041	0.0436	0.8649	0.2700
Predicted pecuniary prepayment probability	0.3499	0.3613	0.1658	0.1692	-0.0114	-0.0680	0.9599	-0.4200
Predicted non-pecuniary prepayment probability	0.3432	0.3387	0.1500	0.1391	0.0045	0.0309	1.1642	0.1900
Original LTV	87.0267	86.6800	11.3053	11.2588	0.3467	0.0307	1.0083	0.1900
FICO score at origination	685.3600	686.8400	56.5178	58.3772	-1.4800	-0.0258	0.9373	-0.1600
Original loan amount (in \$10,000)	10.7981	12.6511	6.2685	5.9166	-1.8530	-0.3040	1.1225	-1.8600
Full income documentation (0,1)	0.5867	0.4133	0.4957	0.4957	0.1733	0.3520	1.0000	2.1400
Prepayment penalty (0,1)	0.4133	0.3600	0.4957	0.4832	0.0533	0.1097	1.0525	0.6700
Property owner occupied (0,1)	0.6933	0.6267	0.4642	0.4869	0.0667	0.1411	0.9088	0.8600
Property condo (0,1)	0.1867	0.4933	0.3923	0.5033	-0.3067	-0.6842	0.6074	-4.1600
Neighborhood-level past housing price appreciation rate at origination	0.1747	0.2716	0.1660	0.2830	-0.0969	-0.4178	0.3439	-2.5600
Neighborhood-level heterogeneity in housing price at origination (in \$10,000)	6.4645	10.1242	5.6529	6.7166	-3.6597	-0.5896	0.7083	-3.6100
Neighborhood-level housing occupancy rate at origination	0.9136	0.8799	0.0491	0.0794	0.0337	0.5105	0.3824	3.1300
Neighborhood-level poverty rate at origination	0.2195	0.1991	0.1279	0.1020	0.0204	0.1762	1.5723	1.0800
Neighborhood-level median income at origination (HMDA)	0.8665	1.0887	0.3086	0.3194	-0.2222	-0.7075	0.9334	-4.3300
Number of matched pairs	75							

^a Matching Sample 1 between African American borrowers and non-Hispanic White borrowers is generated with a matching caliper radius ratio of 0.8.

^b The standardized difference of the mean, the ratio of the variance, and mean difference *t*-test of all the variables in loan contract rate determination equation between the two racial (ethnic) groups are used for matching balance diagnosis following the matching literature (Rubin, 2001; Austin, 2011). Following Rubin (2001), a variable is well balanced if and only if the standardized difference of the mean falls in the range of (-0.25, 0.25), and the ratio of the variance falls in the range of (0.5, 2).

Table B.3 Matching Balance Diagnosis on Matching Sample 2 between African Americans and Non-Hispanic Whites ^a

Variables	Mean		Std. Dev.		Mean diff.	Std. mean diff. ^b	Ratio of the var. ^b	Mean diff. <i>t</i> -test ^b
	African American	Non-Hispanic White	African American	Non-Hispanic White				
10-year treasury yield at origination (y_0)	4.9467	4.9243	0.7628	0.8362	0.0225	0.0281	0.8322	0.1900
Predicted default probability	0.1021	0.0904	0.0920	0.0940	0.0118	0.1264	0.9568	0.8400
Predicted pecuniary prepayment probability	0.3452	0.3554	0.1673	0.1698	-0.0102	-0.0604	0.9703	-0.4000
Predicted non-pecuniary prepayment probability	0.3499	0.3535	0.1486	0.1416	-0.0036	-0.0246	1.1002	-0.1600
Original LTV	88.1011	88.1685	11.2058	10.7716	-0.0674	-0.0061	1.0822	-0.0400
FICO score at origination	685.0449	688.7191	56.5207	56.0604	-3.6742	-0.0653	1.0165	-0.4400
Original loan amount (in \$10,000)	11.7862	13.6070	7.5057	6.0257	-1.8208	-0.2675	1.5516	-1.7800
Full income documentation (0,1)	0.6067	0.4157	0.4912	0.4956	0.1910	0.3893	0.9823	2.5800
Prepayment penalty (0,1)	0.4494	0.4157	0.5003	0.4956	0.0337	0.0681	1.0187	0.4500
Property owner occupied (0,1)	0.7303	0.7191	0.4463	0.4520	0.0112	0.0252	0.9750	0.1700
Property condo (0,1)	0.1685	0.4157	0.3765	0.4956	-0.2472	-0.5648	0.5769	-3.7500
Neighborhood-level past housing price appreciation rate at origination	0.1737	0.2258	0.1585	0.2350	-0.0521	-0.2602	0.4548	-1.7400
Neighborhood-level heterogeneity in housing price at origination (in \$10,000)	6.5897	9.9209	5.7830	6.6341	-3.3311	-0.5353	0.7599	-3.5700
Neighborhood-level housing occupancy rate at origination	0.9166	0.8897	0.0495	0.0855	0.0268	0.3837	0.3344	2.5600
Neighborhood-level poverty rate at origination	0.2150	0.1798	0.1236	0.0994	0.0352	0.3135	1.5472	2.0900
Neighborhood-level median income at origination (HMDA)	0.8744	1.0885	0.3123	0.3370	-0.2141	-0.6589	0.8586	-4.4000
Number of matched pairs	89							

^a Matching Sample 2 between African American borrowers and non-Hispanic White borrowers is generated with a matching caliper radius ratio of 1.0.

^b The standardized difference of the mean, the ratio of the variance, and mean difference *t*-test of all the variables in loan contract rate determination equation between the two racial (ethnic) groups are used for matching balance diagnosis following the matching literature (Rubin, 2001; Austin, 2011). Following Rubin (2001), a variable is well balanced if and only if the standardized difference of the mean falls in the range of (-0.25, 0.25), and the ratio of the variance falls in the range of (0.5, 2).

Table B.4 Matching Balance Diagnosis on Matching Sample 1 between Hispanics and Non-Hispanic Whites ^a

Variables	Mean		Std. Dev.		Mean diff.	Std. mean diff. ^b	Ratio of the var. ^b	Mean diff. t-test ^b
	Hispanic	Non-Hispanic White	Hispanic	Non-Hispanic White				
10-year treasury yield at origination (y_0)	5.1802	5.1847	0.7499	0.7760	-0.0046	-0.0060	0.9337	-0.0700
Predicted default probability	0.0584	0.0625	0.0957	0.1015	-0.0041	-0.0413	0.8893	-0.4500
Predicted pecuniary prepayment probability	0.4353	0.4357	0.1723	0.1719	-0.0004	-0.0022	1.0041	-0.0200
Predicted non-pecuniary prepayment probability	0.3102	0.3144	0.1334	0.1340	-0.0042	-0.0317	0.9919	-0.3500
Original LTV	81.0041	81.2531	11.9288	11.1829	-0.2490	-0.0215	1.1379	-0.2400
FICO score at origination	710.5851	713.1784	45.6594	48.0131	-2.5934	-0.0554	0.9044	-0.6100
Original loan amount (in \$10,000)	18.4297	18.0898	11.5887	11.6267	0.3400	0.0293	0.9935	0.3200
Full income documentation (0,1)	0.3983	0.4481	0.4906	0.4983	-0.0498	-0.1007	0.9691	-1.1100
Prepayment penalty (0,1)	0.2033	0.2116	0.4033	0.4093	-0.0083	-0.0204	0.9709	-0.2200
Property owner occupied (0,1)	0.8133	0.7718	0.3905	0.4206	0.0415	0.1023	0.8622	1.1200
Property condo (0,1)	0.2656	0.3900	0.4426	0.4888	-0.1245	-0.2670	0.8198	-2.9300
Neighborhood-level past housing price appreciation rate at origination	0.1433	0.1695	0.1402	0.2344	-0.0262	-0.1357	0.3578	-1.4900
Neighborhood-level heterogeneity in housing price at origination (in \$10,000)	11.5340	14.5625	8.3745	8.2417	-3.0284	-0.3645	1.0325	-4.0000
Neighborhood-level housing occupancy rate at origination	0.9244	0.8755	0.0829	0.0992	0.0490	0.5355	0.6983	5.8800
Neighborhood-level poverty rate at origination	0.1194	0.1415	0.0677	0.0820	-0.0221	-0.2937	0.6811	-3.2200
Neighborhood-level median income at origination (HMDA)	1.3210	1.4406	0.6042	0.6402	-0.1195	-0.1920	0.8909	-2.1100
Number of matched pairs	241							

^a Matching Sample 1 between Hispanic borrowers and non-Hispanic White borrowers is generated with a matching caliper radius ratio of 0.6.

^b The standardized difference of the mean, the ratio of the variance, and mean difference t-test of all the variables in loan contract rate determination equation between the two racial (ethnic) groups are used for matching balance diagnosis following the matching literature (Rubin, 2001; Austin, 2011). Following Rubin (2001), a variable is well balanced if and only if the standardized difference of the mean falls in the range of (-0.25, 0.25), and the ratio of the variance falls in the range of (0.5, 2).

Table B.5 Matching Balance Diagnosis on Matching Sample 2 between Hispanics and Non-Hispanic Whites ^a

Variables	Mean		Std. Dev.		Mean diff.	Std. mean diff. ^b	Ratio of the var. ^b	Mean diff. <i>t</i> -test ^b
	Hispanic	Non-Hispanic White	Hispanic	Non-Hispanic White				
10-year treasury yield at origination (y_0)	5.1406	5.1342	0.7331	0.7820	0.0064	0.0084	0.8788	0.1000
Predicted default probability	0.0701	0.0696	0.1155	0.1107	0.0005	0.0045	1.0892	0.0500
Predicted pecuniary prepayment probability	0.4250	0.4241	0.1746	0.1751	0.0009	0.0051	0.9946	0.0600
Predicted non-pecuniary prepayment probability	0.3146	0.3198	0.1363	0.1374	-0.0052	-0.0382	0.9845	-0.4500
Original LTV	81.5548	81.3816	11.1654	11.1795	0.1731	0.0155	0.9975	0.1800
FICO score at origination	708.6254	712.5830	44.0976	49.9347	-3.9576	-0.0840	0.7799	-1.0000
Original loan amount (in \$10,000)	19.8664	19.8803	12.2761	12.7310	-0.0139	-0.0011	0.9298	-0.0100
Full income documentation (0,1)	0.3922	0.4558	0.4891	0.4989	-0.0636	-0.1290	0.9610	-1.5300
Prepayment penalty (0,1)	0.2049	0.2226	0.4044	0.4167	-0.0177	-0.0431	0.9416	-0.5100
Property owner occupied (0,1)	0.8198	0.7633	0.3850	0.4258	0.0565	0.1395	0.8176	1.6600
Property condo (0,1)	0.2438	0.3604	0.4301	0.4810	-0.1166	-0.2560	0.7998	-3.0400
Neighborhood-level past housing price appreciation rate at origination	0.1594	0.1775	0.1840	0.2315	-0.0181	-0.0864	0.6316	-1.0300
Neighborhood-level heterogeneity in housing price at origination (in \$10,000)	11.7360	15.0218	8.3624	8.2864	-3.2858	-0.3947	1.0184	-4.7000
Neighborhood-level housing occupancy rate at origination	0.9238	0.8761	0.0801	0.1013	0.0476	0.5216	0.6247	6.2000
Neighborhood-level poverty rate at origination	0.1199	0.1362	0.0676	0.0836	-0.0163	-0.2147	0.6550	-2.5500
Neighborhood-level median income at origination (HMDA)	1.3119	1.4760	0.6055	0.6510	-0.1642	-0.2611	0.8651	-3.1100
Number of matched pairs	283							

^a Matching Sample 2 between Hispanic borrowers and non-Hispanic White borrowers is generated with a matching caliper radius ratio of 0.8.

^b The standardized difference of the mean, the ratio of the variance, and mean difference *t*-test of all the variables in loan contract rate determination equation between the two racial (ethnic) groups are used for matching balance diagnosis following the matching literature (Rubin, 2001; Austin, 2011). Following Rubin (2001), a variable is well balanced if and only if the standardized difference of the mean falls in the range of (-0.25, 0.25), and the ratio of the variance falls in the range of (0.5, 2).

Table B.6 Matching Balance Diagnosis on Matching Sample 3 between Hispanics and Non-Hispanic Whites ^a

Variables	Mean		Std. Dev.		Mean diff.	Std. mean diff. ^b	Ratio of the var. ^b	Mean diff. <i>t</i> -test ^b
	Hispanic	Non-Hispanic White	Hispanic	Non-Hispanic White				
10-year treasury yield at origination (y_0)	5.1412	5.1278	0.7419	0.7824	0.0134	0.0176	0.8992	0.2100
Predicted default probability	0.0710	0.0716	0.1159	0.1156	-0.0006	-0.0055	1.0040	-0.0700
Predicted pecuniary prepayment probability	0.4264	0.4218	0.1751	0.1763	0.0046	0.0262	0.9860	0.3200
Predicted non-pecuniary prepayment probability	0.3137	0.3205	0.1364	0.1376	-0.0067	-0.0492	0.9820	-0.6000
Original LTV	81.3188	81.2651	11.3906	11.0411	0.0537	0.0048	1.0643	0.0600
FICO score at origination	706.4698	712.1510	43.2222	50.5697	-5.6812	-0.1208	0.7305	-1.4700
Original loan amount (in \$10,000)	20.6591	20.9714	12.4177	13.7308	-0.3123	-0.0239	0.8179	-0.2900
Full income documentation (0,1)	0.3926	0.4664	0.4892	0.4997	-0.0738	-0.1496	0.9582	-1.8200
Prepayment penalty (0,1)	0.1913	0.2181	0.3940	0.4137	-0.0268	-0.0666	0.9070	-0.8100
Property owner occupied (0,1)	0.8188	0.7685	0.3858	0.4225	0.0503	0.1246	0.8339	1.5200
Property condo (0,1)	0.2383	0.3557	0.4267	0.4795	-0.1174	-0.2592	0.7919	-3.1600
Neighborhood-level past housing price appreciation rate at origination	0.1638	0.1771	0.1990	0.2284	-0.0134	-0.0624	0.7593	-0.7600
Neighborhood-level heterogeneity in housing price at origination (in \$10,000)	11.7560	15.5081	8.0657	8.4694	-3.7521	-0.4537	0.9069	-5.5400
Neighborhood-level housing occupancy rate at origination	0.9241	0.8748	0.0846	0.1024	0.0493	0.5250	0.6831	6.4100
Neighborhood-level poverty rate at origination	0.1179	0.1338	0.0658	0.0828	-0.0159	-0.2119	0.6322	-2.5900
Neighborhood-level median income at origination (HMDA)	1.3198	1.5194	0.5871	0.6909	-0.1996	-0.3114	0.7221	-3.8000
Number of matched pairs	298							

^a Matching Sample 1 between Hispanic borrowers and non-Hispanic White borrowers is generated with a matching caliper radius ratio of 1.0.

^b The standardized difference of the mean, the ratio of the variance, and mean difference *t*-test of all the variables in loan contract rate determination equation between the two racial (ethnic) groups are used for matching balance diagnosis following the matching literature (Rubin, 2001; Austin, 2011). Following Rubin (2001), a variable is well balanced if and only if the standardized difference of the mean falls in the range of (-0.25, 0.25), and the ratio of the variance falls in the range of (0.5, 2).

CHAPTER 4

CONCLUSIONS

This dissertation research attempts to examine if a borrower's personal traits including gender and race impact loan contract rate beyond the extent to which they affect loan default probability and prepayment probability. In order to solve the potential problem which is common in prior studies that a borrower's personal traits might be associated with loan termination patterns, loan default probability and prepayment probability are modelled via a competing-risks loan hazard model, and are fully controlled for when examining the relationship between a borrower's personal traits and loan contract rate.

The results on loan hazard evaluation indicate that a borrower's personal traits are indeed associated with loan default probability and prepayment probability. Specifically, the results in the first essay on gender equality reveal that female sole borrowers tend to be more likely to default than borrowers in other gender groups, and the results in the second essay on racial equality report that African American borrowers and Hispanic borrowers appear to prepay less frequently than non-Hispanic White borrowers. These results indicate the importance of accounting for the association between a borrower's personal traits and loan termination patterns which is normally ignored in prior studies. Both the regression results and the matching results on the relationship between a borrower's personal traits and loan contract rate provide empirical evidence of gender inequality and racial inequality in mortgage lending. The results in the first essay show female sole borrowers appear to pay significantly higher contracts rates than borrowers in other gender groups, and this gender disparity in contract rate could not be explained by the impact of gender on

tendencies to default or prepay a loan, but rather gender itself. The results in the second essay also provide empirical evidence of adverse pricing against African American borrowers and Hispanic borrowers, as these two racial (ethnic) groups of borrowers are shown to pay significantly higher contract rates than non-Hispanic White counterparts. Also, these contract rate disparities across racial (ethnic) groups are only attributable to race (ethnicity) itself.

Given the observed phenomenon of inequality found in this dissertation research, several possible explanations are proposed. In the first essay, possible sources of gender inequality include discrimination against females by lenders in the mortgage markets, less intensive search efforts by females, less knowledge on the mortgage market by females, and weaker bargaining power by female borrowers. Due to the limitation of the data, those possible sources could not be tested and identified. However, in the second essay, two possible sources of racial (ethnic) inequality – discrimination and regulation, could be identified. The results indicate that the higher contract rates paid by Hispanic borrowers relative to non-Hispanic White borrowers are entirely attributable to regulation, while the higher contract rates paid by African American borrowers relative to non-Hispanic White borrowers are partially because of regulation, and also partially due to discrimination.