The Secret Bank Settlement: Who Was Guilty and Who Was Not?

A Dissertation presented

by

E Yang

to

The Graduate School

in Partial Fulfillment of the

Requirements

for the Degree of

Doctor of Philosophy

in

Economics

(Concentration - Applied microeconomics, Financial economics)

Stony Brook University

August 2015

Stony Brook University

The Graduate School

E Yang

We, the dissertation committe for the above candidate for the

Doctor of Philosophy degree, hereby recommend

acceptance of this dissertation

Warren Sanderson - Dissertation Advisor
Professor, Department of Economics

Sandro Brusco - Dissertation Co-Advisor
Professor and Chair, Department of Economics

Mark Montgomery - Dissertation Examining Committee Chair
Professor, Department of Economics

Alexis Anagnostopoulos - Committee Member
Assistant Professor, Department of Economics

Keli (Andrew) Xiao - Outside Committee Member
Assistant Professor, College of Business

This dissertation is accepted by the Graduate School

Charles Taber
Dean of the Graduate School

Abstract of the Dissertation

Title of Dissertation

by

E Yang

Doctor of Philosophy

in

Economics

(Concentration - Applied microeconomics, Financial economics)

Stony Brook University

2015

Since the subprime mortgage crisis, many banks agreed to pay huge sums to settle the government's accusations that they sold flawed mortgage securities in the 2008 crisis. The Bank of America, JP Morgan, Citi, Wells Fargo, and many other banks have paid a total of more than $130 billion for claims that they intentionally misled investors or were guilty of financial wrongdoing. Is there evidence to support this position and is it fair to assign equal blame to all the big banks? Here, we estimate the expected default rate based on loan characteristics reported by big banks and compare with their actual default rates. We find that, in general, big banks did worse than their predicted loan default rates based on their reported loan characteristics. Loans by the Bank of America and Countrywide not only had an extremely higher default rate than expected, they were also much worse than loans by other big banks and the base group banks. The data also shows that Wells Fargo did better than predicted for most of the years and also did better than the small banks. Our analysis supports the evidence that there should be different levels of settlement with the big banks.

Dedication Page

I dedicate this thesis to my daughter, Anny Yang Chen. You motivated me to complete this process and also sacrificed immensely along the way. I love you!

TABLE OF CONTENTS

LIST OF FIGURES 7

LIST OF TABLES 8

ACKNOWLEDGEMENTS 10

CHAPTER

I. Summary 1

II. Introduction 5

2.1 The Current Situation 8

III. History of Fines 12

3.1 Recent Merger Events 12
3.2 History of Bank Fines and Reasons 14
3.2.1 Bank of America 14
3.2.2 J.P.Morgan Chase 17
3.2.3 Citigroup 17
3.2.4 Wells Fargo 18

IV. History of Fannie and Freddie Policies 20

4.1 The Community Reinvestment Act of 1977 20
4.2 National Homeownership Strategy in 1995 23
4.3 Policy Changes in the Last 15 Years 25
4.3.1 Documentation Relief 28
4.3.2 The Bush Administration and Mortgage Loans Policy 31

V. Data & Traditional Models 35

5.1 Data 35
5.1.1 Evidence of Policy Change 39
5.1.2 What Changed? 44
5.1.3 Missing Data 52
5.2 Literature Review and Traditional Theory 53
5.2.1 Broad Review about Risk Factors 53
5.2.2 Paper from Models Perspective 55
5.2.3 Traditional Binary Predictive Model 57

VI. The Theory of Survival Analysis 62

6.1 Parametric Models 63
6.1.1 Maximum Likelihood 63
6.1.2 Accelerated Failure Method 65
6.1.3 Proportional Hazard Method 66
6.2 Non-Parametric Method 68
6.2.1 The Kaplan-Meier Estimator 69
6.2.2 The Nelson-Aalen Estimator 72
6.2.3 Testing the Equality of Survival Functions 73
6.3 Semi-Parametric Model:Cox Proportional Model 75
6.3.1 Maximum of Partial Likelihood 77
6.3.2 Cox Model Diagnostics 81
6.4 Competing Risk Model 85
6.4.1 Nonparametric Analysis 86
6.4.2 Semiparametric Analysis 87

VII. Application of Survival Analysis 89

7.1 Nonparametric Survival Function 90
7.2 Applying the Cox Proportion Model 93
7.3 Applying a Competing Risk Model 96

VIII. Policy Implications 105

8.1 Default Rate over Time 105
8.2 The Number of Loans Originated over Time 106
8.3 Final Thoughts 107

BIBLIOGRAPHY 109

LIST OF FIGURES

Figure

2.1 Historical Foreclosure Activity 9

3.1 Recent Banks Merge Events 13

5.1 Default Comparison Over time 52

5.2 Missing Data Proportion over Time 52

5.3 Default Comparison across Banks 53

5.4 Missing Data Proportion for Banks 53

5.5 ROC Curve Example-1 60

5.6 ROC Curve Example-2 60

7.1 Kaplan Meier and Nelson Aalen Survivor Function 91

7.2 K-M Survivor Function by Bank 92

7.3 Hazard Rate Cox Proportional 95

7.4 Hazard Rate Nonparametric 95

7.5 Cox-Snell Residual vs. NA cumulative Hazard 96

7.6 Cumulative Incidence by Bank-FG Method 99

8.1 Default Rate for Loans in Each Quarter of Calendar Years 105

LIST OF TABLES

Table

5.1 Market Share of Banks from 1999 to 2008 35

5.2 The Number of Loans by Banks and States from 1999 to 2008 . . . 38

5.3 State Distribution of Loans Originated from 1999 to 2008 by Big Banks 40

5.4 The Number of Loans Originated by Banks over Time 41

5.5 The Percentage of Loans Originated by Big Banks over Time 41

5.6 Annual Default Rate for Loans Originated in Sample Years 41

5.7 Cumulative Default Rate for Loans Originated in Sample Years . . 42

5.8 Kolmogorov Smirnov Statistic for Credit Scores between Banks . . 45

5.9 Kolmogorov Smirnov Statistic for CLTV between Banks 45

5.10 Kolmogorov Smirnov Statistic for DTI between Banks 46

5.11 Quantile Regression for Three Loan Characteristics across Banks . 49

5.12 Quantile Regression for Three Loan Characteristics over Years . . . 50

5.13 Contingency Table 59

6.1 Combination of Model and Specifications 68

6.2 Comparison between PH and AFT for Two Parametric Models . . 68

6.3 Survival Time and Default Status for Five Loans 69

6.4 KM Survival Function Estimator Example 71

6.5 Nelson-Aalen Cumulative Hazard Estimator Example 72

6.6 Kaplan-Meier and Nelson-Aalen Estimator Comparison 73

6.7 Test of Equality of Survivor Function in Two Groups 74

6.8 Test of Equality of Survivor Function 75

6.9 Cox Proportional Example Data - without Tie of Failures 78

6.10 Cox Proportional Example Data - with Tie of Failures 79

7.1 Loans Originated in Florida from 2000-2008 89

7.2 Kaplan-Meier Survivor and Nelson-Aalen Cumulative Hazard Function 90

7.3 Survivor Function by Bank 92

7.4 Log-rank Equality of Survivor Test 93

7.5 Survivor Function Equality Test 93

7.6 Survivor Function by Bank 94

7.7 Results from Cox Model 94

7.8 Cox Model Test 95

7.9 Estimated Results from Six Models 98

7.10 Estimated X-year Cumulative Default Rate for Loans in Florida Originated During 2007-2008 99

7.11 Loan Overall Situation vs SettlementS 100

8.1 The Number of Loans Originated over Time 106

ACKNOWLEDGEMENTS

First and foremost, I would like to express my sincere gratitude to my advisor, Professor Warren Sanderson, for the continuous support of my Ph.D study and research, for his inspiration, enthusiasm, immense knowledge and encouragement. His guidance helped me through all the time of research and writing of this thesis. Without his advising, it would have been impossible for me to finish this thesis, from idea to model structure, from motivation to conclusion. I will be grateful for all of my life.

Besides this, I would like to sincerely thank Professor Sandro Brusco: his critical decisions completely change my track and make me closer to my dream. I would also thank Professor Mark Montgomery and Professor Hugo Benitez-Silva for their encouragement and insightful comments. I also thank the rest of my thesis committee: Dr. Alexis Anagnostopoulos, and Dr. Keli Xiao, for their support and great questions. My sincere thanks also go to Professor Thomas Muench, Dr. Samuele Centorrino, Dr. Yiyi Zhou and many other faculty members who attended seminar and gave me suggestions. I also owe thanks to Professor Patrick Campbell at Long Island University for warm encouragement and careful proofreading.

I thank all of my friends at Stony Brook University: Yajun Wang, Bo He, Yan Yuan, Ling Zhang, Lunan Jiang, Siyu Qi, Jinyoung Eom, Yi Zhang, and Yuanyuan Deng, for the classes we took together, for the projects we worked on together, for all the fun we have had in the last few years. Last, but not the least, I would like to thank my parents, Xiuyun Xu and Yanchang Yang, for bringing me to the world in the first place, and thank my husband, Lei Chen, for spiritually supporting me all the time.

CHAPTER I

Summary

In the mortgage loan market, financial institutions (banks and other mortgage loan institutions) approve loans applications and sell the loans to the Federal National Mortgage Association (FNMA, also called Fannie Mae) and the Federal Home Loan Mortgage Corporation (FHLMC, also called Freddie Mac), the nation's two largest government-sponsored mortgage investors.

Freddie Mac securities carry no government guarantee of being repaid. Although it is explicitly stated in the securities themselves, as well as in public communications issued by Freddie Mac, there is widespread belief that Freddie Mac securities are backed by some sort of implied federal guarantee and a majority of investors believe that the government would prevent a disastrous default. Eventually, overconfidence in the market with other factors led to a financial crisis, and then banks settled with huge amount of money with the government for their wrongdoings.

While the banks have paid clearly for their actions, neither the details of the settlements have been made public nor are the reasons why the decision was not made public. Many are left to wonder whether or not the big banks really did something wrong, since the regulators have not been able to bring out the details to the public despite the record fines the big banks have paid. Secret political deals like these undermine investors' confidence in markets by decreasing the predictability and clarity of the law; this is not only bad for the law, but also bad for the capitalist system.

In this thesis, I apply the survival analysis model to the Freddie Mac 30-year fixed rate loans and compare different performance of loans approved by big banks, including Chase, Citi, Bank of America, Countrywide, Wells Fargo, with the loans approved by smaller banks.

We noticed that banks have different strategies in assessing mortgage loans: some banks require a high credit score, while other banks care more about a high standard with LTV or DTI ratios. The loans approved by different banks have different characteristics in terms of the distribution of credit score, DTI, and LTV ratio. And for the same bank, it changed its policy over the years. It is not surprising to see the different banks have different loan default rates, and that the default rate for the same bank changed over the years, providing the fact that the loan characteristics are different across banks and over time and the real estate market changes over time.

We also noticed that different banks have different market share in each state of the U.S, and different state suffered from the financial crisis differently. To make the default rate across banks comparable, we focused on the mortgage loans approved in the same state, and compared the default rate across banks for loans that were originated during the same time period.

From the model results, we see that when loan characteristics are controlled for loans those originated in the financial crisis period, all big banks, except Wells Fargo, have positive and significant coefficients, which indicate that they have a higher probability of hazard than loans by smaller banks, when loans characteristics are controlled. Wells Fargo had a negative and significant coefficient, which indicates that Wells Fargo did a better job than small banks in terms of loan management efficiency.

From the comparison between the actual and predicted loan default rate, we can see how banks are different in the efficiency level of mortgage loans management. Bank of America and Countrywide ranked at the bottom; Citi and Chase are slightly better than small banks. Wells Fargo did better than small banks.

In addition, we discuss about the missing data in the data section. Although only 2.5% of the data is missing, it provides a good perspective as to whether there was a problem. In general, the loans with missing data should be low credit loans, and are expected to default more frequently than loans with complete data. As was pointed out earlier, Bank of America and Countrywide have a lower default probability for loans with missing data than loans with complete data. This is inconsistent with other banks and with our intuition.

Provided with the different number/proportion of loans approved by different banks during the financial crisis, we can see that there is solid ground for a different level of settlement for each bank.

Overall, there were significant reasons for big banks to agree to the huge settlements. Although the information is not publicly available, we have found evidence from the public data that Bank of America and Countrywide did misrepresent the quality of their loans, from both missing data and non-missing data loan performance. The loans by Bank of America and Countrywide have the highest hazard rate over time and the lowest survival rate over time, when all other factors are included. Bank of America and Countrywide also performed much worse than their predicted default rate during the financial crisis. This explains that $74 billion settlement, the largest penalty levied against all of the big banks with the federal government. Part of the reason Bank of America and Countrywide did poorly is the good work they did in earlier years, which led to their later over-confidence in underestimating loan risk, especially when the market turned around. This does not justify their wrongdoing, since the loans they originated performed much worse than predicted.

Chase and Citi performed worse than expected, but not worse than other banks, which explains why their settlements are lower than Bank of America's. Citi and Chase did poorly in the earlier years and were more cautious in making later loans, so did not fare as badly. According to the difference between the actual predicted default rates, Citi's situation was slightly worse than Chase's, however, during the financial crisis period, Citi approved 77,793 loans, while Chase approved 182,878 loans. With very close loan performance, the number of loans plays a more important role in the settlement decision.

Wells Fargo did the best of all the big banks, the coefficients in all models are significantly better, and this is not coincidental. For Wells Fargo, the number of loans underwritten during the financial crisis is 367,205, which seems large, but it is only 11.8% of all loans approved since 1999. For Bank of America, their percentage of loans approved during financial crisis is 30.2%; for Countrywide, their loans approved during the financial crisis period accounts for 38.7% of all their loans. Hence, Wells Fargo was more cautious about approving loans.

Above all, regulators made reasonable settlement deals with most banks; although from this perspective, justice can be based on extortion. While Wells Fargo did the best of all the big banks, it was not accidental, since during 2001-2005, Wells Fargo was continuously accused (and punished) for number of alleged wrongdoing by the SEC and other regulators. Had they been guilty of illegal activities, they would have had to pay fines, which is perhaps why they kept a cautious eye on all their later loans, which performed much better than the average in the industry. If a bank like Wells Fargo could perform well during the financial crisis, and still get punished, then we might well need to rethink about what justice means in a free society. Wells Fargo continued to battle FHA to avoid further punishment as late as May 2014, and failed. Although it failed in June, this sends a message that if the bank believes in its innocence, and does not deserve such severe punitive measures, it will not give up the legal battle or easily settle for with such an unfair agreement. For now we can only wait and see what the final outcome will be.

CHAPTER II

Introduction

The mortgage market shares the same characteristics across the world; buying a home with a mortgage represents a leveraged investment in which owners can accumulate equity. As long as there is leverage, there is also the risk of default. While banks do not bear default risk, they do not make money on interest. Banks do charge origination fees, service fees, and charge fees when selling mortgage loans. Although they originate loans, banks usually cannot afford to keep all the loans they provide. The Federal National Mortgage Association (FNMA, also called Fannie Mae) and Federal Home Loan Mortgage Corporation (FHLMC, also called Freddie Mac), the nation's two largest government-sponsored mortgage investors, purchase mortgage loans from banks, package similar loans and then sell them as mortgage-backed securities.

As of 2008, Fannie Mae and Freddie Mac owned or "guaranteed" about half of the U.S.'s $12 trillion mortgage market. This made both corporations highly susceptible to the subprime mortgage crisis of that year. Ultimately, in July 2008, the speculation became a reality when the US government took action to prevent the collapse of both corporations. On September 7, 2008, the U.S. Government took control of both Fannie Mae and Freddie Mac.

In October of 2010, government estimates revealed that the bailout of Freddie Mac and Fannie Mae would likely cost taxpayers $154 billion. Taxpayers finally did foot the bill, but this was far from the end. In August 2014, Bank of America agreed to pay $16.65 billion to settle the government's accusations it sold flawed mortgage securities in the 2008 crisis, the largest settlement ever reached between the U.S. and a single company. This adds up to BOA's 19th settlement, for a total payout of $74.58 billion. With JP Morgan's 8 settlements amounting to $27.09 billion, Citibank's $12.14 billion in 8 settlements, Wells Fargo's $9.9 billion in 9 settlements, Morgan Stanley's $1.91 billion, Suntrust's $1.1 billion, and GMAC's $1.2 billion, the nation's largest banks have paid close to $130 billion for supposedly misleading investors in mortgage-backed bonds.

While the banks have paid a costly price for their actions, neither the details of the settlements have been made public nor are the reasons why the decision was not made public. Many are left to wonder whether or not the big banks really did something wrong, since the regulators have not been able to bring out the details to the public despite the record fines the big banks have paid. Secret political deals like these undermine investors' confidence in markets by undermining the predictability and clarity of the law; this is not only bad for law, but also bad for the capitalist system.

Senator Elizabeth Warren and Senator Tom Coburn have put forward a bill to make the terms of such settlements public in the future. A recent article (cri (2014)) in *the Economist* says that the heads of big firms chose to pay such large corporate penalties in order to avoid personal criminal charges, even though they are innocent. In this paper, we ask the question of whether there is evidence that Countrywide (taken over by Bank of America in 2008) and other banks misrepresented risk on mortgage loans.

We also will discuss whether there is reasonable support for different levels of penalties against different banks. A high default rate does not necessarily mean the banks misrepresented the loans, if the loans underwritten by Countrywide or other banks had low average credit scores, high debt to income ratios and high loan to value ratios. In other words, high risk loans are expected to have a higher default rate.

To arrive at this conclusion, we first estimated the expected default rate for each bank, and then compared the predicted default rate with the actual default rate for each bank. We find that almost all the big banks performed worse than their expected default rate during financial crisis period. Bank of America had a predicted 6-year default rate lower than Citi and Wells Fargo; however, Bank of America had an actual default rate of much greater than Citi and Wells Fargo's actual default rate. Countrywide had an actual 6-year default rate of 19.94%, or 45% greater than its predicted default rate (13.74%).

Second, we focus on how the loan performance by the big banks compared with other banks (the base group) by a semi-parametric model, parametric models, and competing risk models. To see the effect of loans during the financial crisis period, we focused on the loans that originated during 2004-2008. The results show that during the financial crisis, the performance of loans by Chase and Citi seemed better than the base group, but not significantly; the loans by Bank of America and Countrywide perform much worse than the base group banks; Wells Fargo did much better than the base group as well as other big banks.

Fannie Mae and Freddie Mac have completely recovered; they have even repaid U.S. taxpayers $6.8 billion after reporting third-quarter profits that modestly rose from the second quarter. According to news, once they have made their most recent payments in December, the two companies will have returned $225.5 billion to taxpayers in exchange for about $188 billion in taxpayer aid that they received after being placed under government protection at the height of the financial crisis.

At the time banks were bashed by politicians for their reckless mortgage lending, and they paid a high price for the lesson. But did regulators learn anything from this? It appears that they have not. According to economists (cri (2014)), Fannie Mae and

Freddie Mac are structurally unsound. In a speech on October 20, Mel Watt, head of the Federal Housing Finance Authority (FHFA), announced plans to reintroduce mortgages with deposits as low as 3% through Fannie Mae and Freddie Mac, the two government-backed housing giants it regulates.

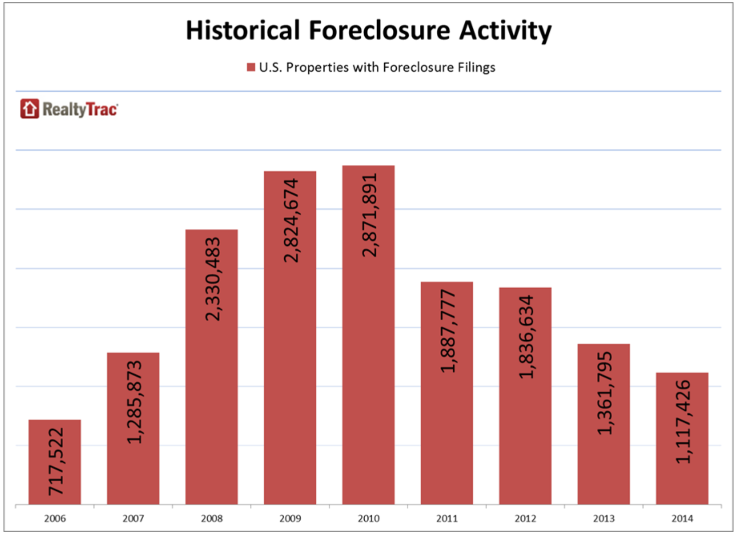
In this paper, we argue that while regulators were correct in distinguishing bad banks from good banks, regulators failed to do an adequate job in taking responsibility themselves for what went wrong. From the public data, we find evidence of Freddie Mac misrepresenting to investors the quality of their loans in their public material. Freddie admitted that they had some high risk loans, but also emphasized that they had bought bad loans at low cost, implying they were clear with potential inventors about the relationship between price and risk.

2.1 The Current Situation

According to RealtyTrac 2014 U.S. Foreclosure Market Report, foreclosure filings default notices, scheduled auctions and bank repossessions were reported on 1,117,426 U.S. properties in 2014, down 18 percent from 2013 and down 61 percent from the peak of 2,871,891 properties with foreclosure filings in 2010. The 1.1 million properties with foreclosure filings in 2014 was the lowest annual total since 2006, when there were 717,522 properties with foreclosure filings nationwide. The report also shows that 0.85 percent of all U.S. housing units (one in every 118) had at least one foreclosure filing in 2014, the first time since 2006 that the annual foreclosure rate has been below 1 percent of all housing units(*Irvine* (2015)).

Figure 2.1 shows that the U.S. foreclosure numbers from 2006 to 2014 resemble the foreclosure market in 2004 and is close to finding a floor and stabilizing at a historically normal level. Does this mean that we have left the 2008 financial crisis and have nothing to worry about? On the one hand, it is true that foreclosure events have gradually decreased. On the other hand, we still have a relatively high number

Figure 2.1: Historical Foreclosure Activity



source:RealtyTrac

of cumulative house foreclosures. How fast we can get out of a bad situation also depends on the sales of these foreclosures.

Take Long Island for example. According to local news, Long Island communities are littered with empty, neglected homes – from small Cape Cod-style houses in Levittown, America's first suburb, to large colonials in upscale Hampton communities. Neighborhoods across Long Island are battling an epidemic of blighted, abandoned houses. Thousands of homes went into New York's nearly three-year-long foreclosure process, creating what have become known as "zombie houses." These homes, with no owner on site and no one taking care of the property, are neither dead nor alive. The abandoned houses ruin the quality of life for neighbors, threaten public safety and send property values plummeting.

The size and scope of the abandoned-home scourge is growing so fast that it challenges municipal efforts to keep up with it. According to data from municipalities and RealtyTrac, a national real estate tracking company, as of January 31 2015, Suffolk County had 2,084 zombie homes and Nassau had 1,960 ranking them seventh and ninth highest, respectively, among 2,165 counties in the United States, the most recent figures available. Suffolk and Nassau are the top counties in New York State for zombie homes; Long Island has the top five ZIP codes in the state for the number of zombie homes.

Local municipalities last year spent at least $3.2 million to clean, maintain, board up and demolish homes in disrepair, including zombie properties. Zombie houses have cost Long Island at least $295 million in depreciated home values, according to a real estate appraiser's analysis.

Bank officials said the homeowner – whether still living on the property or not – is legally responsible for the continued maintenance through the foreclosure process. The financial institutions aren't responsible because they don't own the properties until a foreclosure judgment is issued.

But government officials and residents say the financial companies should do more to protect the properties from deteriorating. In thousands of cases, Long Island municipalities have stepped in to ensure public safety and protect property values, using tax dollars and their own work crews to clean up, board up and tear down deteriorating abandoned houses. Even more public employee time is spent trying to find the property owner or the bank to take action.

According to *Bonilla* (2015),"As of Jan. 31, there were 182 properties considered zombie houses in the Bay Shore ZIP code, according to data from California-based RealtyTrac, which identifies zombie homes through county foreclosure records and postal service information. Hempstead Village had 170; Brentwood 168; Freeport 142; and Central Islip 139. Few communities are spared. In Suffolk, Holbrook had 42 zombie homes and East Northport had 27. In Nassau, Westbury had 79, Hicksville 49 and Glen Cove 36. Five are located in the East Hampton ZIP code. Westhampton Beach officials on Thursday demolished the boarded-up, crumbling home of incarcerated former Suffolk Legis. George Guldi. The house had been gutted by fire in 2008 and deteriorating since then."

"Brookhaven spent more than $800,000 cleaning, boarding up or

> tearing down blighted homes in 2014. In Islip, the town spent more than $200,000 on abandoned homes in 2014, while Hempstead spent more than $700,000 on residential properties that fell into disrepair. Municipalities undertake work at an abandoned house for a variety of reasons: houses neglected by absentee owners; properties where ownership is in dispute; homes being rehabilitated but which have run out of financing; and zombie homes in foreclosure."

The remainder of this paper is organized as follows: Chapter III and IV briefly introduce the history of fine and mortgage loan policies. Chapter V describes the data and introduce the traditional models. Chapter VI and VII introduce the survival analysis theory and apply the survival analysis tools to the data. Chapter VIII discusses mortgage loan policy implication, pointing out some systematic risk we may not be able to avoid.

CHAPTER III

History of Fines

3.1 Recent Merger Events

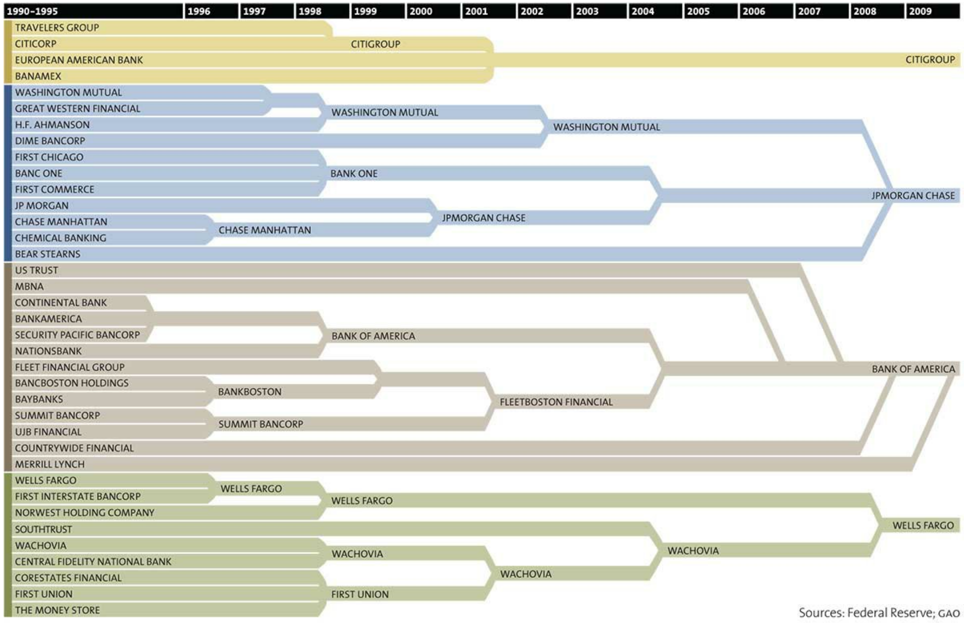
The history of the financial industry is a history of about merger and acquisitions; from 1996 to 2008, there were 74 banks mergers and acquisitions. The merger and acquisitions by Bank of America, Citi, Chase, and Wells Fargo accounted for about 40% of the total.

Here we want to mention Countrywide and Wachovia to help us understand why some banks were accused and fined for wrongdoing by other banks.

After American Home Mortgage Investment Corporation filed bankruptcy in August,2007, attention began to focus on Countrywide Financial, which issued 17% of all mortgages in the United States. On August 16, Fitch Ratings downgraded Countrywide Financial Corporation to BBB+, its third lowest investment-grade rating, and Countrywide borrowed the entire $11.5 billion available in its credit lines with other banks. However, this could not prevent investors from worrying about Countrywide's potential bankruptcy risk.

Bank of America deployed many analysts to model the performance of Countrywide's loan portfolio. In January 2008, after several weeks analyzing, Bank of America announced that it planned to purchase Countrywide Financial for $4.1 billion in stock, while the stock market value was $24 billion the year before. In June 2008,

Figure 3.1: Recent Banks Merge Events



Sources: Federal Reserve; GAO

source: Federal Reserve

Bank of America Corporation announced it had received approval from the Board of Governors of the Federal Reserve System to purchase Countrywide Financial Corporation. Countrywide shareholders approved the deal in the same month. On July 2008, Bank of America Corporation completed its purchase of Countrywide Financial Corporation.

The deal was a landmark in the housing crisis, given the fact that Countrywide was the largest lender, issuing 17% of all mortgages in the United States at that time.

Wachovia was another big merge in the housing crisis. Wachovia was the fourth-largest bank holding company in the United States based on total assets before its acquisition by Wells Fargo in 2008. At its peak time, Wachovia was one of the largest providers of financial services in the United States, operating in more than 20 states in the United States and providing global services through more than 40 offices around the world. The purchase of Wachovia by Wells Fargo and Company was completed on December 31, 2008.

3.2 History of Bank Fines and Reasons

According to recent news from *the Financial Times*, a US judge ruled on May 11, 2015 that Nomura and Royal Bank of Scotland had misled investors in mortgage-backed securities, “offering documents that did not correctly describe the mortgage loans” and “The magnitude of falsity, conservatively measured is enormous”.

The judge ruled in favor of a government agency acting on behalf of Fannie Mae and Freddie Mac, whose vast portfolios of mortgage-backed securities plummeted in value during the crisis, jeopardising the companies future and triggering a controversial government bailout.

The Federal Housing Finance Agency, conservator for Fannie and Freddie, sued 19 financial institutions in 2011. Seventeen of these institutions, ranging from JPMorgan Chase to Barclays, have agreed to pay more than $20 billion to settle allegations that they mis-sold the securities to Fannie and Freddie.

According to Bloomberg Business News on February 25, 2015, Morgan Stanley agreed to pay $2.6 billion to settle probes into its creation and sale of residential mortgage-backed securities, as the U.S. Department of Justice holds another large Wall Street firm to account for the 2008 financial crisis. Goldman Sachs Group Inc. disclosed the same week that it received a letter from the U.S. Attorneys Office in Sacramento, saying a civil lawsuit may be brought against the firm.

The negotiations between FreddieMac/FannieMae and the banks is still ongoing.

3.2.1 Bank of America

Bank of America has paid a total about $74 billion in several settlements. In August 2014, Bank of America agreed to pay $16.65 billion, the biggest settlement in history(*Backman* (2014)). In July 2014, a federal judge ordered Bank of America to pay an additional $1.27 billion penalty for fraud over shoddy mortgages sold by the former Countrywide Financial Corp. The case centered on a Countrywide lending program that ended around May 2008, known as "High Speed Swim Lane," "HSSL" or "Hustle." According to Reuters, the program emphasized quantity over quality, rewarding employees for producing more loans and eliminating checkpoints designed to ensure the loans' quality. "While the HSSL process lasted only nine months, it was from start to finish the vehicle for a brazen fraud by the defendants, driven by a hunger for profits and oblivious to the harms thereby visited, not just on the immediate victims but also on the financial system as a whole,"

Earlier in 2014, it paid the Federal Housing Finance Authority more than $9 billion and settled for $1.3 billion with the U.S. Attorney in New York's Southern District; in 2013, it paid Fannie Mae $10 billion more for mortgages and forked over nearly $3 billion more for foreclosures; in 2012, it paid nearly $12 billion to settle lawsuits over wrongful foreclosures and more than $2 billion in a class-action suit it inherited from Merrill Lynch; in 2011, it paid trustee Bank of New York $8.6 billion and bond insurer Assured Guaranty $1.6 billion after they filed lawsuits over bond deals that went sour; in 2010, it paid $2.8 billion to Fannie Mae and Freddie Mac over mortgages.

While many of the mortgage securities in question were made by Countrywide and Merrill Lynch, the government found problems with Bank of America's own mortgage securities as well, including efforts to circumvent underwriting standards by changing applicants' financial information. In at least one instance, an underwriter at Bank of America made more than 40 attempts to win an "accept" rating from an internal Countrywide system, known as CLUES, that would allow Bank of America to make a loan, according to a statement of facts signed by the U.S. and Bank of America (*Rexrode and Grossman* (2014)).

In March 2014, as conservator of Fannie Mae and Freddie Mac, the Federal Housing Finance Agency (FHFA) announced it has reached a settlement in cases involving Bank of America, Countrywide Financial, Merrill Lynch. Countrywide and Merrill

Lynch were taken over by Bank of America in 2008. The cases alleged violations of federal and state securities laws in connection with private-label, residential mortgage-backed securities (PLS) purchased by Fannie Mae and Freddie Mac between 2005 and 2007. Allegations of common law fraud were made in the Countrywide and Merrill Lynch cases. The Agreement provides for an aggregate payment of approximately $9.33 billion by Bank of America that includes the litigation resolution as well as a purchase of securities by Bank of America from Fannie Mae and Freddie Mac

In 2013, according to *Isidore* (2013), Bank of America paid $3.55 billion in cash to Fannie as part of the deal. It also repurchases 30,000 questionable mortgages likely to produce losses, paying Fannie $6.75 billion for the loans. The loans had been bundled into mortgage-backed securities, which then were bought and guaranteed by Fannie Mae.

In 2012, it paid nearly $12 billion to help settle lawsuits over wrongful foreclosures and more than $2 billion in a class-action suit it inherited from Merrill Lynch. In 2011, it paid trustee Bank of New York $8.6 billion and bond insurer Assured Guaranty $1.6 billion after they filed lawsuits over the bond deals that went sour. In 2010, it paid $2.8 billion to Fannie Mae and Freddie Mac over mortgages.

The total bill is a large sum, even for a big bank: It made just over $16 billion in profit from 2010 to 2013 – and about $2 billion more in the past six months. The sum has been inflated by lawsuits against subprime home lender Countrywide Financial, which it bought in 2008, and investment bank Merrill Lynch, which was purchased in 2009.

The amount is more than the combined total of all the other major banks, which have paid out about $56 billion in financial crisis settlements and fines. With the most recent $16.65 billion, Bank of America has paid a total about $74 billion in several settlements.

3.2.2 J.P.Morgan Chase

J.P. Morgan Chase agreed to pay $13 billion in November 2013, including a $4 billion payment for consumer relief, along with a payment to investors of more than $6 billion and a large fine. In September and October 2013, it paid $1 billion to end investigations into the botched financial transactions of traders in London that cost the company more than $6 billion; in September, it refunded $389 million to 2.1 million credit card customers and paid a fine after allegedly misleading and overcharging them, and resolved an insurance lawsuit at the cost of $300 million, splitting payment with Assurant Inc; in August, it paid $410 million to settle allegations that it manipulated U.S. energy markets; in January 2013, 10 banks, including JPMorgan, split a settlement of $8.5 billion related to wrongful home foreclosures. In February 2012, five of the nation's largest banks, including Chase, split a $25 billion settlement over charges of systemic and widespread mortgage fraud, in what is being billed as the largest-ever deal on such charges.

3.2.3 Citigroup

Citigroup agreed to pay $7 billion in July 2014 to settle a U.S. investigation into shoddy mortgage-backed securities that the bank sold in the run-up to the financial crisis, including the largest civil fraud penalty ever levied by the U.S. Justice Department. In January 2013, 10 banks, including Citi, split a settlement of $8.5 billion related to wrongful home foreclosures. In February 2012, five of the nation's largest banks, including Citi, split a $25 billion settlement over charges of systemic and widespread mortgage fraud, in what is being billed as the largest-ever deal on such charges.

3.2.4 Wells Fargo

In November 2009, Wells Fargo agreed to buy back $1.4 billion in auction-rate securities to settle allegations by the California attorney general of misleading investors. In May 2011 it was fined $1 million by FINRA for failing to send disclosure documents to customers. That same month, it agreed to pay up to $16 million to settle charges of violating the Americans with Disabilities Act. In July 2011 Wells Fargo agreed to pay $125 million to settle a lawsuit in which a group of pension funds accused it of misrepresenting the quality of pools of mortgage-related securities. That same month, the Federal Reserve announced an $85 million civil penalty against Wells Fargo for steering customers with good qualifications into costly subprime mortgage loans during the housing boom.

In November 2011 Wells Fargo agreed to pay at least $37 million to settle a lawsuit accusing it of municipal bond bid rigging. The following month, FINRA fined it $2 million for improper sales of reverse convertible securities and later another $2.1 million for failing to properly supervise the sale of exchange-traded funds. Wells Fargo was one of five large mortgage servicers that in February 2012 consented to a $25 billion settlement with the federal government and state attorneys general to resolve allegations of loan servicing and foreclosure abuses. The New York Attorney General later sued Wells Fargo for breaching the terms of that settlement.

In July 2012 the U.S. Justice Department announced that Wells Fargo would pay $175 million to settle charges that it engaged in a pattern of discrimination against African-American and Hispanic borrowers in its mortgage lending during the period from 2004 to 2009. In August 2012 Wells Fargo agreed to pay $6.5 million to settle SEC charges that it failed to fully research the risks associated with mortgage-backed securities before selling them to customers such as municipalities and non-profit organizations.

In October 2012 the U.S. Attorney for the Southern District of New York filed suit against Wells Fargo, charging the bank with engaging in a longstanding practice of reckless underwriting and fraudulent loan certification for thousands of loans insured by the Federal Housing Administration that ultimately defaulted. And in January 2013 Wells Fargo was one of ten major lenders that agreed to pay a total of $8.5 billion to resolve claims of foreclosure abuses.

In June 2013 Wells Fargo settled a lawsuit alleging that it neglected the maintenance and marketing of foreclosed homes in black and Latino areas by agreeing to spend at least $42 million to promote home ownership and neighborhood stabilization.

In October 2013 Freddie Mac announced that Wells Fargo would pay $869 million to repurchase home loans the bank had sold to the mortgage agency that did not conform to the latter's guidelines.

CHAPTER IV

History of Fannie and Freddie Policies

4.1 The Community Reinvestment Act of 1977

There is a long term American policy of promoting home ownership. The earliest important policy we can find is in The Community Reinvestment Act of 1977. According to Wiki, The Community Reinvestment Act is a United States federal law designed to encourage commercial banks and savings associations to help meet the needs of borrowers in all segments of their communities, including low- and moderate-income neighborhoods. Congress passed the Act in 1977 to reduce discriminatory credit practices against low-income neighborhoods, a practice known as redlining. With red lining, banks can refuse to grant mortgages for the the purchase of houses in certain neighborhoods with significant high risks. The risk can be associated with a decline in market value caused by a high crime rate, vandalism, or the emergence of the problem of gangs. It is a rational business decision to avoid granting loans within such neighborhood.

However, when those people with an appropriate income level to repay a mortgage were turned down, and they found similar people with similar loan application in another neighborhood was accepted, they felt the banks were falsely measuring the risk of not having the mortgage repaid. Those who had their mortgage applications turned down could easily assume that the neighborhood the house of the mortgage was associated with had been red lined. It would be easy to attribute the rejection to some socio-economic aberrations. The primary socio-economic aberrations blamed were racial discrimination, because African-Americans and Hispanics are the least successful mortgage applicants.

The mortgage rejections were especially plausible when the unsuccessful applicants had average or above average incomes. While income is only one part of the evaluation process, credit history is equally significant. The banks wants to ascertain whether they will be repaid. A borrower with a bad credit history is less likely to repay a loan than a borrower with a good credit history but lower income. Irresponsible behavior can dissipate even a large income.

Denial of access to credit also implies denied opportunity to receive capital gains. A large number of those unsuccessful mortgage applicants felt it was unfair to deny them access to credit, and they attributed their lesser wealth to not having access to such capital gains. In the 1970's, there was fertile field for community organizers to exploit.

My understanding of this subject was enhanced by reading *Watkins et al.* (2009). According to *Watkins et al.* (2009), one of these community organizers in Chicago was Gale Cincotta, who was convinced that redlining was preventing her neighborhood in west Chicago from developing. She and other members of a community organization talked to bank officials. When that did not produce the results she wanted she organized an action in which her followers interrupted bank business by flooding the bank with demands for time consuming services. Following that action she organized a march of 600 to a meeting of the Chicago City Council. When that action failed to produce results she organized 1200 to make a protest at the Department of Housing and Urban Development, a federal government agency. That action produced significant results, including an investigation of the situation in west Chicago by the Nixon Administration.

Gale Cincotta then sought to prevent a bank from closing its local branch. She organized a protest at the bank and the bank capitulated, keeping its branch open and promising to invest several million dollars in west Chicago. This victory in 1972 led her to found two organizations, the National People's Action (NPA) and the National Training and Information Center (NTIC).

By 1975 she had acquired significant political power and organized a conference in Chicago on the matter of red lining. That brought the issue to the attention of the general public and Congress. In 1976, Congress passed the Home and Mortgage Disclosure Act and Gerald Ford signed it into law. This act required banks to disclose where they were granting mortgages. That information allowed Gale Cincotta and other community organizers to make accusations of red lining and racial discrimination against banks. The media picked up and publicized the issue.

In 1976, Gale Cincotta announced the formulation of new policy that ultimately became the Community Reinvestment Act (CRA). She propounded the notion that it was immoral for banks to take the deposits of people in one community and lend them for investment in another community. Behind this idea was the belief, held by many not in business, that businesses can easily operate at a profit and constraining their actions for social purposes imposes no cost on their operations. According to Gale Cincotta, banks have a duty to lend to people in the neighborhoods where they operate.

Cincotta's ideas caught the attention of William Proxmire, U.S. Senator for Wisconsin. Proxmire's background was in journalism and he saw no problem in forcing banks to lend to people in the neighborhoods where they operate, even if such loans cannot be justified on the basis of profitability. Proxmire believed that something like Cincotta's proposal would be required to end racial discrimination of banks.

Proxmire had his staff cooperate with Cincotta and her associates. As a result, the Community Reinvestment Act of 1977, a piece of legislation of enormous financial and economic consequence, was written by Gale Cincotta and her community organizer associates.

The bill as written by Gale Cincotta and Shel Trapp, another community organizer, required that banks maintain records of where they made loans and that information would be made available to community organizations. As written, the CRA provided that when a bank appeared before regulatory agencies, the comminity organizers had a right to testify about the bank's fulfillment of its duty to serve the needs of the community in which it operates. This enabled community organizers to extort large donations from banks. If a bank wanted to undertake any new action it knew that it would have to payoff the community organizers to get the request approved by the bank regulators.

From the Homestead Act in 1862 to the GI Bill of Rights in 1944, there is nothing wrong in helping people realize the American dream. However, to promote homeownership, we need to be more careful with the economy situation because the mortgage market uses leverage, a 10:1 leverage can magnified the loss or profit 10 times. When loans are originated, buyers buy with leverage and when mortgage-backed securities are sold in the market, hedge funds use leverage. In a bad macro economy, more people lost their jobs, driving them out of the home. More inventory on the market push the housing price down, which leads to investors loss in the bond market which is backed by mortgage loans. With less liquidity provided by financial institutions, demand becomes weaker, and pushes down the housing price, and so on.

4.2 National Homeownership Strategy in 1995

As we have mentioned, promoting home ownership is a long term policy. One of the critical documents we should mention is"The National Homeownership Strategy: Partners in the American Dream" published by U.S. Department of Housing and Urban Development (HUD) in 1995. According to the report's background introduction, in the spring and summer of 1994, Secretary Henry Cisneros met with leaders of major national organizations from the housing industry to solicit their views about establishing a national home ownership partnership. In August 1994, these planning sessions culminated in a historic meeting at which industry representatives agreed to the formation of working groups to help develop the National Homeownership Strategy. Hundreds of people from more than 50 organizations met frequently from late August through mid-December, and eventually developed this plan in May 1995.

In the President's Message, the first part of the report, President Clinton announced the goal of adding as many as eight million new families to America's home ownership rolls by the year 2000. This report identifies specific actions that the federal government, its partners in state and local government, the private, nonprofit community, and private industry would take to lower barriers that prevented American families from becoming homeowners. The report listed 51 actions, and one of the action (Action 37) even suggested using IRAs and 401ks for Home ownership Downpayments. HUD analysis indicates that at least 600,000 households in the next 5 years would benefit from withdrawing funds from their retirement accounts for a first-time downpayment option.

President George W. Bush continued promoting homeownership. In a speech to HUD employees in June 2002, he brought up the homeownership gap in America. "Three-quarters of Anglos own their homes, and yet less than 50 percent of African Americans and Hispanics own homes. That ownership gap signals that something might be wrong in the land of plenty. And we need to do something about it." "We are here in Washington, D.C. to address problems. So I've set this goal for the country. We want 5.5 million more homeowners by 2010."

It is believed in the National Home Ownership Strategy that home ownership creates economic prosperity for families and communities and acts as a dynamic generator of economic growth. Every new home creates 2.1 jobs directly related to construction, and many more jobs through increased demand for household goods and services. Based on this, it is not understand why our successive presidential administrations have promoted home ownership.

According to the HUD/FHA Annual Management Report (2006), two successes in this year were mentioned. One was support for first - time homebuyers, 79.3 percent of FHA-insured purchase loans involved first-time homebuyers, providing 248,953 families the ability to purchase their first home. The other was to assist homeowners facing financial difficulties remain in their homes. FHA again encouraged lenders to increase their use of loss mitigation tools. As a result, loss mitigation cases increased from 24,874 cases in Fiscal Year 1999 to 75528 in Fiscal Year 2006, an increase of 204%. However, even as the situation worsened, HUD/FHA completely ignored this signal, "FHA has proposed legislation that would revitalized the federal government's largest mortgage program. The bill, which overwhelmingly passed in the House 415-7 and awaits Senate action, would allow the FHA to offer flexible down payment options for the first time, increase permissible mortgage amounts substantially in high-cost markets, and provide low-interest rates and consumer protections that are rarely available from 'sub-prime' mortgage lenders.The FHA would join the rest of mortgage market in underwriting home buyers based on their risk of default as measured by credit scores, down-payment amounts and financial profiles, thus allowing more lower-income borrowers the possibility of enjoying the many benefits FHA offers. "

4.3 Policy Changes in the Last 15 Years

Both the National Homeownership Strategy and the Community Reinvestment are too broad and to cover a 40 year history, it is hard to see their affect directly. Fannie Mae and and Freddie Mac are the nation's biggest underwriters of home mortgages. Although they don't lend money directly into consumers, they purchase loans that banks make on the secondary market. Their extending or contracting the loans buying rules can affect the banks and small mortgage underwriting institutions. Now we follow how their policies in have evolved over the last 15 years.

Holmes (1999) reports that Fannie Mae eased credit to aid mortgage lending in 1999. According to this report, the Fannie Mae Corporation eased the credit requirements on loans that it purchased from banks and other lenders, which would help increase home ownership rates among minorities and low-income consumers. The action, which began as a pilot program involving 24 banks in 15 markets – including the New York metropolitan region – will encourage those banks to extend home mortgages to individuals whose credit is generally not good enough to qualify for conventional loans. Fannie Mae officials hoped to make it a nationwide program by 2000 spring.

According to *Holmes* (1999), "Under Fannie Mae's pilot program, consumers who qualify can secure a mortgage with an interest rate one percentage point above that of a conventional, 30-year fixed rate mortgage of less than $240,000 – a rate that currently averages about 7.76 per cent. If the borrower makes his or her monthly payments on time for two years, the one percentage point premium is dropped. Fannie Mae officials stress that the new mortgages will be extended to all potential borrowers who can qualify for a mortgage. But they add that the move is intended in part to increase the number of minority and low income home owners who tend to have worse credit ratings than non-Hispanic whites. Home ownership has, in fact, exploded among minorities during the economic boom of the 1990's. The number of mortgages extended to Hispanic applicants jumped by 87.2 per cent from 1993 to 1998, according to Harvard University's Joint Center for Housing Studies. During that same period the number of African Americans who got mortgages to buy a home increased by 71.9 per cent and the number of Asian Americans by 46.3 per cent. In contrast, the number of non-Hispanic whites who received loans for homes increased by 31.2 per cent."

As the nation's biggest underwriter of home mortgages, Fannie Mae was pressured both from the Clinton Administration and stock holders. Government hopes to expand mortgage loans among low and moderate income people, while stock holders hoped to maintain its phenomenal growth in profits. Moreover, banks and mortgage companies had been pressing Fannie Mae to help them make more loans to so-called subprime borrowers. These borrowers whose incomes, credit ratings and savings were not good enough to qualify for conventional loans, could only get loans from finance companies that charge much higher interest rates – anywhere from three to four percentage points higher than conventional loans.

"Fannie Mae has expanded home ownership for millions of families in the 1990's by reducing down payment requirements," said Franklin D. Raines, Fannie Mae's chairman and chief executive officer. "Yet there remain too many borrowers whose credit is just a notch below what our underwriting has required who have been relegated to paying significantly higher mortgage rates in the so-called subprime market."

In *Holmes* (1999), it was pointed out directly that a bad economy might bring government-subsidized corporations into trouble. "In moving, even tentatively, into this new area of lending, Fannie Mae is taking on significantly more risk, which may not pose any difficulties during flush economic times. But the government-subsidized corporation may run into trouble in an economic downturn, prompting a government rescue similar to that of the savings and loan industry in the 1980's."

The failure of Fannie Mae was also predicted in this report by Peter Wallison, a resident fellow at the American Enterprise Institute: "From the perspective of many people, including me, this is another thrift industry growing up around us. If they fail, the government will have to step up and bail them out the way it stepped up and bailed out the thrift industry."

> "Despite these gains, home ownership rates for minorities continue to lag behind non-Hispanic whites, in part because blacks and Hispan-

> ics in particular tend to have on average worse credit ratings. In July, the Department of Housing and Urban Development proposed that by the year 2001, 50 percent of Fannie Mae's and Freddie Mac's portfolio be made up of loans to low and moderate-income borrowers. Last year, 44 percent of the loans Fannie Mae purchased were from these groups. The change in policy also comes at the same time that HUD is investigating allegations of racial discrimination in the automated underwriting systems used by Fannie Mae and Freddie Mac to determine the credit-worthiness of credit application."

4.3.1 Documentation Relief

Freddie Mae and Freddie Mac continued easing credit through offering "reduced documentation" or "no documentation", also called "documentation relief". It was explained in a famous blog by *Dungey* (2007), who worked as loan underwriter for mortgage institutions, that reduced documentation loans were originated in two general ways, lender-directed or borrower-directed. Lender-directed means that the lender first looks at the loan as a whole. If the applicant has a good credit history, lenders will offer reduced income documentation, adding up the total amount of the deposits for 3 or 6 months and then dividing that total by 3 or 6 months, and using this amount for the applicant's average monthly income.

Many banks do this because they believe that the loan applicants may have other income that isn't documented. Undocumented income can be rental income, a side business, or any income from loans to family or friends. Self-employed borrowers and cash tip earners typically prefer reduced documentation loans, since people usually reduce their reported income for tax purpose. The reduced documentation loan reduces the amount of paperwork and eliminates many steps required when applying for a loan. Compared with full-documentation loans, income needs to be verified for the last two years, while reduced documentation loans can be quite selective and biased. Even for a salaried borrower, the average monthly income in the last 3 months can vary a lot from the average monthly income from the last two years. However, when the applicant has a good credit history or the loan looks good, the lender might require the loan applicant to submit only the last pay stub, instead of requiring W-2s for the last two years.

Lender-directed program, criticized by *Dungey* (2007), is not really just "documentation relief," it is also a way to approve a loan application with a marginally higher income than it would have been with full documentation loan. In general, salary workers are assumed to have an up trend in salary, it is not a big problem, however, it can be a major risk when commissioned borrowers, contract workers are considered.

The borrower-directed program is much riskier than lender-directed program. In the borrower-directed program, the borrower requests a reduced documentation or no documentation loan, which causes serious problems when using an Automatic Underwriting System (AUS) like Loan Prospector (LP) used by Freddie Mac and Desktop Underwriter (DU) used by Frannie Mae to underwrite these loans. LP and DU might allow some documentation relief after the initial analysis is done, but all documentation relief is based on the assumption that any information the applicant provides for income or assets is verifiable. Obviously, the borrower can always inflate their income or asset on purpose, then the system categorize their loans into a certain group so that they qualify for reduced documentation loans. The ridiculous part is that the borrower always can get the AUS sytem come up with a "documentation relief" offer if they lie in the first place.

Although now we only have access to the Guidelines of Mortgage Underwriting for the time after 2010 from Freddie Mac official web. It is confirmed in *Fitch* (2010) that reduced documentation was widely used. "In the height of the housing boom in 2006 and 2007, reduced documentation loans accounted for roughly 40% of newly issued mortgages in the U.S., according to mortgage-data firm FirstAmerican CoreLogic. University of Chicago assistant professor Amit Seru says that for subprime loans, the portion exceeded 50%."

It is also mentioned that the most outrageous types of no-doc lending disappeared entirely in 2009. Many mortgage pros say they are unaware of banks making any low-doc loans in recent months. In fact, the financial reform package passed by the House of Representatives recently, and under consideration by the Senate, discourages them from doing so. It requires lenders who offer mortgages to borrowers without full documentation to post a reserve equal to 5% of the loan's value before they are securitized. That rule, they say, will make low-doc loans even less appealing for banks going forward.

In 2008, *Setzer* (2008) reports that Fannie Mae Tightens Loan Criteria for Credit Scores. "Fannie Mae's Managing Director, Brian Faith, released a statement on Wednesday that gave notice that at least one of the two government sponsored enterprises (GSEs) that play a major role in the nation's mortgage industry has decided it would be wise to protect its own interests"

In 2010, people were concerned that cautious about reduced documentation loan was back, *Fitch* (2010) reported that"Wall Street Funding of America, a mortgage lender based in Santa Ana, Calif., was recently circulating offers to make low-doc loans to borrowers with credit scores as low as 660 on the Fair Isaac Corp. (FICO) scale, as long as the borrower was self-employed, seeking no more than 60% of the value of a home and had six months of mortgage payments in reserve. The lender was offering interest rates 1.5 to 2 percentage points over the going rate on conventional mortgages. A borrower with a credit score over 720 might get a slightly better rate, perhaps just 1.25 percentage points over. On June 23 Wall Street Fundings fliers caught the attention of Zillow.com blogger Justin McHood. Forbes calls to Wall

Street Funding were not returned. "

4.3.2 The Bush Administration and Mortgage Loans Policy

On April 12, 2000, it was reported that Fannie Mae Moves Against Predatory Loans: "The nation's largest home loan financier announced guidelines to fight predatory practices in the exploding mortgage market for buyers with low incomes and poor credit. In a letter to mortgage lenders, Fannie Mae–a public company chartered by the government that buys one of four home loans made by banks–said it would refuse to purchase most mortgages with upfront fees of more than 5% of the loan amount.The agency said it will also deny most loans with prepayment penalties and with credit life insurance policies.The steps follow increasing concerns expressed by government officials and community groups in recent weeks about the so-called subprime lending industry, which typically makes loans that carry higher fees and interest rates to home buyers with low incomes and poor credit."

On June 12, 2002, the HUD chief urged Fannie Mae and Freddie Mac to work harder for minorities, help low-income people and minorities buy homes. The Bush administration has yet to describe its policy toward the shareholder-owned but congressionally chartered companies, known as government-sponsored enterprises, or GSEs, because of the benefits Congress grants them to help promote home ownership. But in the past two weeks, White House and Treasury officials have issued statements calling for tighter regulation of GSE financial disclosure and boards of directors and expressing concern that retail investors may mistakenly think GSE securities are guaranteed by the government. Martinez said the GSEs may not have lived up to their congressionally mandated goal of outperforming the broader market in financing home loans for segments of the population that tend to be less likely to own their own homes.

On June 17 2002, at St. Paul AME Church in Atlanta, Georgia, president Bush calls for expanding opportunities to home ownership, saying "And part of economic security is owning your own home. Part of being a secure America is to encourage homeownership. So somebody can say, this is my home, welcome to my home. Now, we've got a problem here in America that we have to address. Too many American families, too many minorities do not own a home. There is a home ownership gap in America. The difference between Anglo America and African American and Hispanic home ownership is too big. And we've got to focus the attention on this nation to address this."

"And it starts with setting a goal. And so by the year 2010, we must increase minority home owners by at least 5.5 million. In order to close the homeownership gap, we've got to set a big goal for America, and focus our attention and resources on that goal."

On June 18, 2002, at HUD Washington, DC, President George W. Bush Spoke to HUD Employees to celebrate National Homeownership Month "I'm here to celebrate National Homeownership Month, because I believe owning a home is an essential part of economic security. And I'm concerned about the security of America. "

Oct. 15, 2002, at George Washington University, President George W. Bush, President Hosts Conference on Minority Homeownership: "We can put light where there's darkness, and hope where there's despondency in this country. And part of it is working together as a nation to encourage folks to own their own home."

On December 16, 2003, Remarks on Signing the American Dream Downpayment Act:"One of the biggest hurdles to home ownership is getting money for a downpayment. This administration has recognized that, and so today I'm honored to be here to sign a law that will help many low-income buyers to overcome that hurdle and to achieve an important part of the American Dream. The law I sign today will help us build on this progress in a very practical way. Many people are able to afford a monthly mortgage payment but are unable to make the downpayment, and so this legislation will authorize $200 million per year in downpayment assistance to at least

40,000 low-income families. These funds will help American families achieve their goals and, at the same time, strengthen our communities." "This Administration will constantly strive to promote an ownership society in America. We want more people owning their own home. It is in our national interest that more people own their own home. After all, if you own your own home, you have a vital stake in the future of our country."

Feb 21, 2005, Bushs 2nd Inaugural Address: "We will widen the ownership of homes and businesses, retirement savings and health insurance, preparing our people for the challenges of life in a free society." Note that while people are living and breathing and existing in America today, they have to be "prepared" according to Mr. Bush for the challenges of life in a free society. Implicit in the sentence is that private ownership of homes, businesses, and private (non-Social Security) retirement accounts are a preparation for "life in a free society." His two statements imply that by acquiring property, citizens will be preparing themselves to live in a free society.

In 2003, the Bush Administration sought to create a new agency, replacing the Office of Federal Housing Enterprise Oversight, to oversee Fannie Mae and Freddie Mac. In 2005, the Federal Housing Enterprise Regulatory Reform Act, sponsored by Senator Chuck Hagel (R-NE) and co-sponsored by Senators Elizabeth Dole (R-NC), John McCain (R-AZ) and John Sununu (R-NH), would have increased government oversight of loans given by Fannie Mae and Freddie Mac. Like the 2003 bill, it also died in the Senate Banking, Housing, and Urban Affairs Committee, this time in the 109th Congress. A full and accurate record of the congressional attempts to regulate the housing GSEs is given in the Congressional record prepared in 2005.

The Housing and Economic Recovery Act of 2008 passed by the United States Congress on July 24, 2008 with bipartisan support and signed into law by President George W. Bush on July 30, 2008. This enabled expanded regulatory authority over Fannie Mae and Freddie Mac by the newly established FHFA, and gave the U.S.

Treasury the authority to advance funds for the purpose of stabilizing Fannie Mae or Freddie Mac, limited only by the amount of debt that the entire federal government is permitted by law to commit to. The law raised the Treasury's debt ceiling by $800 billion, to a total of $10.7 trillion, in anticipation of the potential need for the Treasury to have the flexibility to support Fannie Mae, Freddie Mac, or the Federal Home Loan Banks.

CHAPTER V

Data & Traditional Models

5.1 Data

The data we use in this paper is the Single Family Loan-Level Dataset recently made public by Freddie Mac. The data set only includes 30 year fixed rate mortgage loans originated from January 1, 1999, through December 31, 2012, with monthly loan performance data through June 30, 2013, which were either sold to Freddie Mac or issued in Freddie Mac Participation Certificates. The data includes approximately 16 million loans from 1999 to 2012, and the performance status from the reported year to 2013.

Table 5.1: Market Share of Banks from 1999 to 2008

Bank	Total Loan Number	Proportion of Total	Note
Chase	772,360	6.23%	
Citi	615,491	4.97%	
Bank of America	537,210	4.34%	
Countrywide	613,681	4.95%	
WellsFargo	3,122,603	25.20%	
Fifththird	165,799	1.34%	
USbank	399,510	3.22%	
BBT	183,568	1.48%	
Flagstar	152,331	1.23%	
ABN	1,301,321	10.50%	merged by Citi
Oldkent	69,039	0.56%	merged by Fifththird
TaylorR,Bean&Whitaker	237,943	1.92%	merged by BoA
Norwest	238,921	1.93%	merged by WellsFargo
Principalresidential	459,069	3.70 %	merged by Citi
Washingtonmutual	319,672	2.58 %	merged by Chase
All others	3,203,702	25.96%	
Total	12,392,220	100%	

Source: Freddie Mac Single Family Loan-Level Dataset

Table 5.1 shows the share of banks for 12.39 millions loans originated from 1999 to 2008. From the table, we can see that Chase underwrote 6.23% of the total loans, Citi, Bank of America, Countrywide take about 5% separately, and Wells Fargo approved 25% of the total loans, Fifth Third, U.S.bank, BBT, Flagstar took 1.34%, 3.22%, 1.48%, 1.23% respectively. Loans by ABN took 10.5%, but ABN Amro Mortgage was purchased by Citigroup in early 2007. All other banks are either small, or merged with other big banks later on, so we will focus on these five big banks.

All the loans in the data set are fully amortizing and categorized as "full documentation." In each year, loans are separated into four files according to which quarter they were reported to Freddie Mac. There were two files for loans reported in each quarter from 1999 to 2012.

One includes loan characteristics, and the other is about loan performance for each month after it was originated. In the loan characteristics file are included 25 variables: credit score, first payment date, first time homebuyer flag, maturity date, metropolitan statistical area, mortgage insurance percentage, number of units, occupancy status, original combined loan to value ratio, original debt to income ratio, original unpaid balance, original loan to value ratio, original interest rate, channel (indicates whether a broker or correspondent), prepayment penalty mortgage, product type (all 30-year fixed rate), property state, property type, the Metropolitan Statistical Area, loan sequence number, loan purpose, original loan term, number of borrowers, seller's name, and service name.

The other file includes the performance status for each loan in every month. After origination, the loan age and the remaining months are reported each month. If a loan's balance was reduced to zero, then the effective date is reported and a code indicating the reason why it is marked to zero. When a loan delinquency reaches 180 days (D180), it will be marked as zero balance. There are several other reasons that cause a zero balance, such as: prepayment (voluntary payout), third party sale prior to D180, short sale or short payoff prior to D180, deed-in-lieu of foreclosure prior to D180, repurchase prior to D180, and real estate owned (REO) acquisition prior to D180. Except for a short sale and a repurchase prior to D180, all other events imply an equivalent event of delinquency of 180 days, which we define as a default in this paper.

If the credit score is less than 301 or greater than 850, the score will be disclosed as "unknown," which will be indicated by three blank spaces. The DTI will be disclosed with actual value when it is from 0% up to 65%, and disclosed as "unknown" when it is greater than 65% and with a null value when it is unknown. If the original LTV ratio is less than 6% or greater than 105%, the ratio will be disclosed as "unknown," which will be indicated by a blank space in the loan record. If the CLTV is less than 0% or greater than 200%, or is less than the original LTV, or the original LTV is "unknown", the ratio will be disclosed as "unknown". Based on this, loans with unknown data is with are low quality and imply a higher default rate than loans with all known data. 327,504 loans out of total 12,719,724 loans are with more than one important loan characteristics missing. After dropping about 2.5% missing loans, we have total loans 12.39 million.

Table 5.2 shows the number of loans in each state by bank. Each column shows the the number of loans in each state originated by different banks. From the left to the right, they are small banks, Chase, Citi, Bank of America, Countrywide, and Wells Fargo. The table shows that most big banks provide loans for most of the states, and California has the largest number of loans, and Florida state has the second highest number of loans. Different banks have a different number of loans in each state.

To see this more clearly, we calculated the proportion of loans in each state for all banks. Only if all the banks had the same loan distribution over different states, did we analyze the overall data and compare across banks. Table 5.3 shows the percentage of loans in each state for all banks. The distribution is different for banks.

Table 5.2: The Number of Loans by Banks and States from 1999 to 2008

State	Base Group	Chase	Citi	Bank of America	Countrywide	Wells Fargo	Total
AK	11,239	253	142	176	1,093	16,562	29,465
AL	96,326	7,572	2,448	1,228	11,370	22,393	141,337
AR	27,186	5,262	2,031	3,237	4,730	18,877	61,323
AZ	158,681	20,556	27,390	16,884	25,347	116,148	365,006
CA	604,616	67,005	76,073	122,038	83,141	440,599	1,393,472
CO	165,051	14,452	11,566	15,265	19,999	102,224	328,557
CT	72,611	7,182	4,623	5,735	7,591	27,840	125,582
DC	7,609	1,325	2,426	2,807	784	6,101	21,052
DE	22,338	4,339	1,780	982	2,287	15,812	47,538
FL	382,724	99,120	24,913	62,439	33,430	205,368	807,994
GA	231,216	28,624	16,633	17,750	13,737	94,104	402,064
GU	2,040	0	0	1	29	0	2,070
HI	13,257	1,315	1,236	4,411	5,779	11,640	37,638
IA	82,698	2,896	2,311	2,528	4,008	37,505	131,946
ID	39,502	1,916	3,268	1,794	5,811	31,843	84,134
IL	414,314	37,036	39,357	14,832	23,104	123,008	651,651
IN	194,674	10,346	23,225	3,768	10,401	38,820	281,234
KS	64,450	3,911	4,647	4,910	5,156	23,962	107,036
KY	114,764	6,764	9,936	3,515	7,662	24,275	166,916
LA	46,164	14,954	4,455	763	5,628	18,858	90,822
MA	179,669	9,660	14,950	13,065	15,700	49,820	282,864
MD	132,469	23,989	32,196	13,948	14,771	84,195	301,568
ME	33,285	1,368	1,098	1,331	3,059	9,566	49,707
MI	440,097	20,055	26,655	4,438	22,255	50,824	564,324
MN	207,007	6,614	7,998	7,232	12,564	149,967	391,382
MO	186,308	14,801	10,675	13,144	14,492	52,780	292,200
MS	20,636	3,377	1,268	554	3,188	12,063	41,086
MT	20,555	1,152	804	336	3,693	16,811	43,351
NC	219,235	24,259	11,247	23,319	13,817	90,870	382,747
ND	10,586	207	472	460	729	6,934	19,388
NE	49,387	614	1,346	649	2,215	22,314	76,525
NH	50,083	3,514	2,292	1,958	3,790	16,152	77,789
NJ	162,550	26,810	23,707	10,441	19,251	97,448	340,207
NM	39,481	2,980	4,205	3,602	4,903	20,632	75,803
NV	56,398	6,099	6,582	7,965	9,819	47,606	134,469
NY	223,736	57,738	9,578	12,231	14,041	103,718	421,042
OH	336,739	19,456	38,305	6,134	15,129	63,791	479,554
OK	56,685	10,291	2,314	4,509	5,383	18,265	97,447
OR	99,897	10,695	8,161	8,116	15,936	92,587	235,392
PA	209,138	34,138	30,167	8,717	19,310	114,146	415,616
PR	22,583	0	220	0	0	87	22,890
RI	23,508	2,289	2,996	2,219	2,649	7,313	40,974
SC	101,926	15,229	5,952	16,340	6,350	33,973	179,770
SD	12,404	405	541	66	1,237	12,817	27,470
TN	95,300	12,529	5,477	8,185	9,516	40,714	171,721
TX	241,311	61,506	43,870	36,930	31,940	202,484	618,041
UT	82,707	5,275	7,941	1,665	11,857	40,426	149,871
VA	198,966	26,086	30,841	21,024	18,334	83,400	378,651
VI	434	0	0	0	0	0	434
VT	29,799	631	937	1,813	1,173	4,582	38,935
WA	176,734	18,699	13,870	15,228	24,360	108,156	357,047
WI	197,836	11,294	6,900	4,465	15,569	70,173	306,237
WV	27,469	2,954	1,504	489	1,668	5,822	39,906
WY	11,045	482	344	217	2,175	7,901	22,164
total	6,707,423	770,024	613,873	535,853	611,960	3,114,276	12,353,409

For example, 9% of loans approved by small banks are in California, while 22.77% of loans approved by Bank of America are in California; Chase has 8.7% of loans in this state and Wells Fargo has 14.15% of loans in this state. It is observed that different areas suffered from the 2008 financial crisis differently, in other words, housing price dropped differently in different areas, which further affected the default rate differently. If we simply analyze the whole data and compare with different bank groups, it is the same as comparing apples to oranges.

We will face the same problem if we do not distinguish time. Table 5.4 shows the number of loans originated by banks over time. The first column shows the number of loans approved by small banks from 1999 to 2008; the second column is for loans approved by Chase; the third column is for Citi; the fourth is for Bank of America, then Countrywide and Wells Fargo. By comparing each row, we can estimate the market share of banks in each year, and see that the share of small banks keeps decreasing, from about 90% in 1999 to 50% in 2008. In other words, big banks increased their market share of mortgage loans during this period.

To see how each bank develops its market over time, we calculated the proportion of loans in each year for all banks (see Table 5.5). We can see that small banks approved more loans in earlier years than in recent years. Chase and Citi approved a relatively large number of loans during 2003-2005, and then started to tighten policy again. Countrywide approved fewer loans in the early years, but underwrote a large number of loans from 2005. Wells Fargo underwrote most of loans during 2001-2005, and approved fewer loans during the financial crisis period.

5.1.1 Evidence of Policy Change

Table 5.6 shows the annual default rate after the loans were originated for the three sample years. Panel A is for loans that originated in 1999, and Panel B and C are for loans that originated in 2002 and 2006. As is shown in Panel A for loans in

Table 5.3: State Distribution of Loans Originated from 1999 to 2008 by Big Banks

State	Base Group	Chase	Citi	Bank of America	Countrywide	Wells Fargo
AK	0.17%	0.03%	0.02%	0.03%	0.18%	0.53%
AL	1.44%	0.98%	0.40%	0.23%	1.86%	0.72%
AR	0.41%	0.68%	0.33%	0.60%	0.77%	0.61%
AZ	2.37%	2.67%	4.46%	3.15%	4.14%	3.73%
CA	9.01%	8.70%	12.39%	22.77%	13.59%	14.15%
CO	2.46%	1.88%	1.88%	2.85%	3.27%	3.28%
CT	1.08%	0.93%	0.75%	1.07%	1.24%	0.89%
DC	0.11%	0.17%	0.40%	0.52%	0.13%	0.20%
DE	0.33%	0.56%	0.29%	0.18%	0.37%	0.51%
FL	5.71%	12.87%	4.06%	11.65%	5.46%	6.59%
GA	3.45%	3.72%	2.71%	3.31%	2.24%	3.02%
GU	0.03%	0.00%	0.00%	0.00%	0.00%	0.00%
HI	0.20%	0.17%	0.20%	0.82%	0.94%	0.37%
IA	1.23%	0.38%	0.38%	0.47%	0.65%	1.20%
ID	0.59%	0.25%	0.53%	0.33%	0.95%	1.02%
IL	6.18%	4.81%	6.41%	2.77%	3.78%	3.95%
IN	2.90%	1.34%	3.78%	0.70%	1.70%	1.25%
KS	0.96%	0.51%	0.76%	0.92%	0.84%	0.77%
KY	1.71%	0.88%	1.62%	0.66%	1.25%	0.78%
LA	0.69%	1.94%	0.73%	0.14%	0.92%	0.61%
MA	2.68%	1.25%	2.44%	2.44%	2.57%	1.60%
MD	1.97%	3.12%	5.24%	2.60%	2.41%	2.70%
ME	0.50%	0.18%	0.18%	0.25%	0.50%	0.31%
MI	6.56%	2.60%	4.34%	0.83%	3.64%	1.63%
MN	3.09%	0.86%	1.30%	1.35%	2.05%	4.82%
MO	2.78%	1.92%	1.74%	2.45%	2.37%	1.69%
MS	0.31%	0.44%	0.21%	0.10%	0.52%	0.39%
MT	0.31%	0.15%	0.13%	0.06%	0.60%	0.54%
NC	3.27%	3.15%	1.83%	4.35%	2.26%	2.92%
ND	0.16%	0.03%	0.08%	0.09%	0.12%	0.22%
NE	0.74%	0.08%	0.22%	0.12%	0.36%	0.72%
NH	0.75%	0.46%	0.37%	0.37%	0.62%	0.52%
NJ	2.42%	3.48%	3.86%	1.95%	3.15%	3.13%
NM	0.59%	0.39%	0.68%	0.67%	0.80%	0.66%
NV	0.84%	0.79%	1.07%	1.49%	1.60%	1.53%
NY	3.34%	7.50%	1.56%	2.28%	2.29%	3.33%
OH	5.02%	2.53%	6.24%	1.14%	2.47%	2.05%
OK	0.85%	1.34%	0.38%	0.84%	0.88%	0.59%
OR	1.49%	1.39%	1.33%	1.51%	2.60%	2.97%
PA	3.12%	4.43%	4.91%	1.63%	3.16%	3.67%
PR	0.34%	0.00%	0.04%	0.00%	0.00%	0.00%
RI	0.35%	0.30%	0.49%	0.41%	0.43%	0.23%
SC	1.52%	1.98%	0.97%	3.05%	1.04%	1.09%
SD	0.18%	0.05%	0.09%	0.01%	0.20%	0.41%
TN	1.42%	1.63%	0.89%	1.53%	1.56%	1.31%
TX	3.60%	7.99%	7.15%	6.89%	5.22%	6.50%
UT	1.23%	0.69%	1.29%	0.31%	1.94%	1.30%
VA	2.97%	3.39%	5.02%	3.92%	3.00%	2.68%
VI	0.01%	0.00%	0.00%	0.00%	0.00%	0.00%
VT	0.44%	0.08%	0.15%	0.34%	0.19%	0.15%
WA	2.63%	2.43%	2.26%	2.84%	3.98%	3.47%
WI	2.95%	1.47%	1.12%	0.83%	2.54%	2.25%
WV	0.41%	0.38%	0.25%	0.09%	0.27%	0.19%
WY	0.16%	0.06%	0.06%	0.04%	0.36%	0.25%
total	100%	100%	100%	100%	100%	100%

Table 5.4: The Number of Loans Originated by Banks over Time

State	Base Group	Chase	Citi	Bank of America	Countrywide	Wells Fargo	Total
1999	810,916	18,163	16,571	31,479	57,278	0	934,407
2000	416,562	22,601	50,922	69,050	20,662	126,962	706,759
2001	878,291	39,995	93,668	87,400	4,041	388,992	1,492,387
2002	935,116	26,038	81,635	96,056	5,768	547,574	1,692,187
2003	1,036,976	105,517	112,864	24,289	35,084	709,597	2,024,327
2004	494,863	174,029	72,066	34	13,996	353,236	1,108,224
2005	602,067	116,342	76,075	29,849	108,971	328,310	1,261,614
2006	498,979	83,461	32,279	35,867	128,802	292,400	1,071,788
2007	522,768	71,903	25,446	86,148	152,036	199,705	1,058,006
2008	510,885	111,975	52,347	75,681	85,322	167,500	1,003,710
Total	6,707,423	770,024	613,873	535,853	611,960	3,114,276	12,353,409

Table 5.5: The Percentage of Loans Originated by Big Banks over Time

State	Base Group	Chase	Citi	Bank of America	Countrywide	Wells Fargo
1999	12.09%	2.36%	2.70%	5.87%	9.36%	0.00%
2000	6.21%	2.94%	8.30%	12.89%	3.38%	4.08%
2001	13.09%	5.19%	15.26%	16.31%	0.66%	12.49%
2002	13.94%	3.38%	13.30%	17.93%	0.94%	17.58%
2003	15.46%	13.70%	18.39%	4.53%	5.73%	22.79%
2004	7.38%	22.60%	11.74%	0.01%	2.29%	11.34%
2005	8.98%	15.11%	12.39%	5.57%	17.81%	10.54%
2006	7.44%	10.84%	5.26%	6.69%	21.05%	9.39%
2007	7.79%	9.34%	4.15%	16.08%	24.84%	6.41%
2008	7.62%	14.54%	8.53%	14.12%	13.94%	5.38%
total	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%

Table 5.6: Annual Default Rate for Loans Originated in Sample Years

	t=1	t=2	t=3	t=4	t=5	t=6	t=7
Panel A:originated in 1999(N=934,407)							
Chase	0.15%	0.42%	0.45%	0.40%	0.21%	0.19%	0.10%
Citi	0.13%	0.42%	0.55%	0.53%	0.28%	0.25%	0.16%
Bank of America	0.05%	0.18%	0.24%	0.27%	0.17%	0.12%	0.05%
Countrywide	0.05%	0.27%	0.35%	0.42%	0.30%	0.22%	0.15%
Industry Average	0.06%	0.27%	0.37%	0.37%	0.23%	0.16%	0.11%
Panel B:originated in 2002(N=1,692,187)							
Chase	0.17%	0.41%	0.35%	0.30%	0.22%	0.15%	0.21%
Citi	0.33%	0.52%	0.40%	0.31%	0.16%	0.18%	0.26%
Bank of America	0.04%	0.13%	0.12%	0.11%	0.07%	0.11%	0.19%
Countrywide	0.16%	0.35%	0.17%	0.19%	0.12%	0.12%	0.14%
Wells Fargo	0.09%	0.23%	0.21%	0.19%	0.11%	0.10%	0.15%
Industry Average	0.21%	0.39%	0.33%	0.29%	0.18%	0.16%	0.24%
Panel C:originated in 2006(N=1,071,788)							
Chase	0.14%	0.72%	2.12%	3.63%	2.52%	1.45%	0.83%
Citi	0.23%	0.87%	1.80%	3.42%	2.61%	1.54%	1.34%
Bank of America	0.06%	0.42%	2.01%	3.90%	2.74%	1.80%	1.13%
Countrywide	0.19%	1.06%	2.81%	4.66%	3.10%	1.93%	1.28%
Wells Fargo	0.13%	0.57%	1.51%	3.01%	1.98%	1.34%	0.86%
Industry Aerage	0.14%	0.66%	1.79%	3.26%	2.26%	1.45%	0.96%

Source: Freddie Mac Single Family Loan-Level Dataset

1999, Chase and Citi have a default rate that is slightly greater than average, while Bank of America has a default rate lower than average. Countrywide did better than average in the first five years after the loans originated. For loans originated in 2002, Bank of America and Countrywide still did very well, much better than average; loans by Wells Fargo also perform better than average; Citi is slightly worse than average; Chase has a mixed performance, some good and some bad years. For loans that originated in 2006, loan performance completely changed: overall, Chase and Citi were slightly worse than average, while Bank of America and Countrywide have a much higher default rate than average. Loans by Wells Fargo still performed better than average.

In all three panels, we can see a clear trend that the default rate increases over time and then decreases after a certain duration. We need to point out that both Bank of America and Countrywide did very well before the subprime mortgage crisis occured.

Table 5.7: Cumulative Default Rate for Loans Originated in Sample Years

	t<=1	t<=2	t<=3	t<=4	t<=5	t<=6	t<=7
Panel A:originated in 1999 (N=934,407)							
Chase	0.15%	0.58%	1.02%	1.42%	1.64%	1.83%	1.93%
Citi	0.13%	0.55%	1.10%	1.63%	1.91%	2.15%	2.31%
Bank of America	0.05%	0.24%	0.48%	0.75%	0.91%	1.04%	1.08%
Countrywide	0.05%	0.32%	0.67%	1.09%	1.39%	1.61%	1.76%
Industry Average	0.06%	0.33%	0.71%	1.08%	1.31%	1.47%	1.58%
Panel A:originated in 2002(N=1,692,187)							
Chase	0.17%	0.58%	0.92%	1.23%	1.44%	1.59%	1.80%
Citi	0.33%	0.85%	1.25%	1.55%	1.71%	1.89%	2.15%
Bank of America	0.04%	0.17%	0.30%	0.41%	0.48%	0.59%	0.78%
Countrywide	0.16%	0.50%	0.68%	0.87%	0.99%	1.11%	1.25%
Wells Fargo	0.09%	0.31%	0.52%	0.71%	0.82%	0.93%	1.08%
Industry Average	0.21%	0.60%	0.93%	1.22%	1.40%	1.56%	1.80%
Panel B:originated in 2006 (N=1,071,788)							
Chase	0.14%	0.86%	2.98%	6.61%	9.13%	10.58%	11.41%
Citi	0.23%	1.10%	2.90%	6.32%	8.93%	10.47%	11.81%
Bank of America	0.06%	0.48%	2.49%	6.39%	9.13%	10.93%	12.06%
Countrywide	0.19%	1.25%	4.06%	8.72%	11.82%	13.75%	15.03%
Wells Fargo	0.13%	0.70%	2.21%	5.22%	7.20%	8.54%	9.40%
Industry Average	0.14%	0.80%	2.59%	5.84%	8.11%	9.55%	10.52%

Source: Freddie Mac Single Family Loan-Level Dataset

Table 5.7-Panel A shows the cumulative default rate for loans that originated in 1999. The overall cumulative 7-year default rate is 1.58%; Bank of America has a

default rate of 1.08%, lower than the industry average, Countrywide has a 7-year default rate of 1.76%, lower than Chase's and Citi's default rate of 1.93% and 2.31%.

Panel B in Table 5.7 shows that the 7-year cumulative default rate for the whole industry increases from 1.58% in 1999 to 1.80% in 2002; however, loans originated by all four big banks have a lower default rate than in 1999. Chase's 7-year loan default rate decreases from 1.93% to 1.80%; Citi's 7-year default rate decreases from 2.31% to 2.15%; Bank of America's 7-year default rate decreases from 1.08% to 0.78%; and Countrywide's 7-year default rate decreases from 1.76% to 1.08%. Wells Fargo has a 7-year default rate of 1.08% in 2002, far below industry average of 1.80%.

However, this situation changed dramatically in 2006, as is shown in Panel C in Table 5.7. All four big banks perform far worse than the industry as a whole, except for Wells Fargo, which still has a lower default rate than the industry average default rate, and far below the other four big banks. It is also important to point out that the Bank of America and Countrywide loans default rate changes from the the lowest to the highest. Loans approved by Countrywide have the overall highest 7-year default rate of 15.03%, and Bank of America follows, with a 7-year default rate of 12.06%. Chase and Citi have a 7-year default of 11.41% and 11.81%, slightly greater than the industry average default rate of 10.52%.

From Panels A, B and C, it can be seen that the cumulative default rate increases significantly during the crisis; all four big banks changed from better than average in 2002. Take Chase for example: the 5-year default rate increased from 1.44% for loans originated in 2002 to 9.13% for loans that originated in 2006, while the 7-year default rate increased from 1.8% to 11.41%. For Countrywide, this difference is even larger. The 5-year default rate changes from 0.99% to 11.82%; 7-year default rate rises from 1.25% to 15.03%.

5.1.2 What Changed?

The Kolmogorov Smirnov statistic is generally used to test whether a small observation of samples have the same distribution with the whole sample in one dimension. It can also be used to test whether two groups have the same distribution for underlying one-dimensional probability distributions. The empirical distribution function Fn for n iid observations Xi is defined as

$$F_n(x) = \frac{1}{n}\sum_{i=1}^{n} I_{X_i \leq x}$$

where $I_{X_i \leq x}$ is the indicator function, equal to 1 if $Xi \leq x$ and is otherwise equal to 0. The Kolmogorov Smirnov statistic for a given cumulative distribution function F(x) is

$$D_n = \sup_x |F_n(x) - F(x)|$$

where sup x is the supremum of the set of distances. If the sample comes from distribution F(x), then Dn converges to 0 almost certainly in the limit when n goes to infinity. It also can be used to test whether two underlying one dimensional probability distributions differ. In this case, the Kolmogorov Smirnov statistic is

$$D_{n,n'} = \sup_x |F_{1,n}(x) - F_{2,n'}(x)|$$

where $F_{1,n}$ and $F_{2,n'}$ are the empirical distribution functions of the first and the second sample respectively, and sup is the supremum function.

Panels A, B, and C in Table 5.8 show the Kolmogorov Smirnov statistic for credit score for loans originated in 1999, 2002, and 2006 respectively; all of the statistics are significant at the 1% level. For loans that originated in 1999, the statistics change from 0.0307 to 0.0601. This means that the credit score for loans between each big bank is not significant. The statistics in Panel B change from 0.0336 to 0.1494, which

Table 5.8: Kolmogorov Smirnov Statistic for Credit Scores between Banks

	Chase	Citi	BankofAmerica	Countrywide	Wells Fargo
Panel A: KS statistic for loans originated in 1999					
	Chase	Citi	BankofAmerica	Countrywide	
Chase	-	0.0307	0.0328	0.0339	
Citi	0.0307	-	0.0383	0.0601	
BankofAmerica	0.0328	0.0383	-	0.0493	
Countrywide	0.0339	0.0601	0.0493	-	
Panel B: KS statistic for loans originated in 2002					
Chase	-	0.0460	0.1052	0.0781	0.0336
Citi	0.0460	-	0.1494	0.1169	0.0381
BankofAmerica	0.1052	0.1494	-	0.0457	0.1134
Countrywide	0.0781	0.1169	0.0457	-	0.0954
Wells Fargo	0.0336	0.0381	0.1134	0.0954	-
Panel C: KS statistic for loans originated in 2006					
Chase	-	0.1325	0.1057	0.1575	0.0336
Citi	0.1325	-	0.2244	0.0483	0.1060
BankofAmerica	0.1057	0.2244	-	0.2529	0.1295
Countrywide	0.1575	0.0483	0.2529	-	0.1250
Wells Fargo	0.0336	0.1060	0.1295	0.1250	-

Source: Freddie Mac Single Family Loan Data *all significant at 1% level

shows the difference in credit score of loans by banks for 2002 becomes larger than before. Of loans originated in 2002, those by Chase and Wells Fargo have the smallest difference, and loans by Bank of America and Citi have the largest difference in terms of credit score. Similarly, the statistics in Panel C change from 0.0336 to 0.575, which indicates that Chase and Wells Fargo still approved similar loans in terms of credit score: Chase and Countrywide approved most different loans in terms of credit score in 2006.

Table 5.9: Kolmogorov Smirnov Statistic for CLTV between Banks

	Chase	Citi	BankofAmerica	Countrywide	WellsFargo
Panel A: KS statistic for loans originated in 1999					
Chase	-	0.1454	0.0673	0.0770	
Citi	0.1454	-	0.1421	0.2223	
BoA	0.0673	0.1421	-	0.0802	
Countrywide	0.0770	0.2223	0.0802	-	
Panel B: KS statistic for loans originated in 2002					
Chase	-	0.1905	0.0923	0.1224	0.1533
Citi	0.1905	-	0.1343	0.1108	0.0384
BoA	0.0923	0.1343	-	0.0637	0.1226
Countrywide	0.1224	0.1108	0.0637	-	0.0760
Wells Fargo	0.1533	0.0384	0.1226	0.0760	-
Panel C: KS statistic for loans originated in 2006					
Chase	-	0.1251	0.0746	0.0539	0.1041
Citi	0.1251	-	0.0599	0.1174	0.0999
BoA	0.0746	0.0599	-	0.1146	0.0932
Countrywide	0.0539	0.1174	0.1146	-	0.0669
Wells Fargo	0.1041	0.0999	0.0932	0.0669	-

Source: Freddie Mac Single Family Loan Data *all significant at 1% level

Table 5.9 shows the Kolmogorov Smirnov statistic for the combined loan to value ratio for loans that originated in 1999, 2002, and 2006; all of the statistics are significant at the 1% level. Take Panel A, for example, the statistics change from 0.0673 to 0.2223. This signifies that the CLTV (combined loan to value) ratio of loans by Chase originated in 1999 is closest to the CLTV of loans by Bank of America, and is far different from the CLTV of loans by Countrywide. The statistics in Panel B changes from 0.0384 to 0.1905, which shows the difference in CLTV of loans by banks for 2002 is not as big as in 1999. Similarly, the statistics in Panel C change from 0.0599 to 0.1146, which means the difference in CLTV by banks continued shrinking.

Table 5.10: Kolmogorov Smirnov Statistic for DTI between Banks

	Chase	Citi	BankofAmerica	Countrywide	WellsFargo
Panel A: KS statistic for loans originated in 1999					
Chase	-	0.0782	0.1136	0.0306	
Citi	0.0782	-	0.0570	0.0991	
BoA	0.1136	0.0570	-	0.1416	
Countrywide	0.0306	0.0991	0.1416	-	
Panel B: KS statistic for loans originated in 2002					
Chase	-	0.1478	0.0890	0.0763	0.0759
Citi	0.1478	-	0.0680	0.0850	0.0729
BoA	0.0890	0.0680	-	0.0300	0.0378
Countrywide	0.0763	0.0850	0.0300	-	0.0153
Wells Fargo	0.0759	0.0729	0.0378	0.0153	-
Panel C: KS statistic for loans originated in2006					
Chase	-	0.0572	0.0778	0.1095	0.1048
Citi	0.0572	-	0.0523	0.0568	0.0531
BoA	0.0778	0.0523	-	0.0438	0.0358
Countrywide	0.1095	0.0568	0.0438	-	0.0131
Wells Fargo	0.1048	0.0531	0.0358	0.0131	-

Source: Freddie Mac Single Family Loan Data *all significant at 1% level

Table 5.10 shows the Kolmogorov Smirnov statistic for DTI of loans approved by each bank in several years. In 1999, the K-S statistics change from 0.0306 (between Countrywide and Chase)to 0.1416(between Countrywide and Bank of America), which means that loans by Countrywide are most similar to loans by Chase in terms of the debt to income ratio, and loans by Countrywide are most different from Bank of America in terms of debt to income ratio. In 2002, the K-S statistics change from 0.0153 (between Countrywide and Wells Fargo) to 0.1478(between Chase and Citi), which means loans by Countrywide and Wells Fargo are similar in terms of debt to income ratio, and loans by Chase and Citi are most different in terms of debt to income ratio.

The K-S shows the absolute difference between loans by different banks, as shown in three tables, they are significantly different over time and across banks. However, how different are the loans characteristics over time and across banks?

For this, we need to apply quantile regression. The theory of quantile is simple.

$$\hat{q}_\tau = \arg\min_{q \in R} \sum_{i=1}^{n} \rho_\tau |y_i - q|$$

$$= \arg\min_{q \in R} \left[(1-\tau) \sum_{y_i < q} |y_i - q| + \tau \sum_{y_i \geq q} |y_i - q| \right]$$

Suppose the τth conditional quantile function is

$$Q_{Y|X}(\tau) = X\beta_\tau$$

Given the distribution function of Y, β_τ can be obtained by solving

$$\beta_\tau = \arg\min_{\beta \in R^k} E(\rho_\tau |Y - X\beta|)$$

Solving the sample analog gives the estimator of β.

$$\hat{\beta_\tau} = \arg\min_{\beta \in R^k} \sum_{i=1}^{n} (\rho_\tau |Y_i - X\beta|)$$

Table 5.11 shows the quantile regression results for three loan characteristics across banks. Loans that originated in 1999, 2002 and 2006 are shown in Panels A, B and C separately. In each panel, the first three columns are coefficients for credit score at three quartile levels by banks; the middle three columns reflect the coefficients for the combined loan to value ratio; the last three columns reflect the coefficients for debt to income ratio at three different quartile levels. In each panel, five big banks' loan characteristics are compared with the industry average loans. The constant number in the last row shows 25%, 50%, and 75% percentile credit score, combined loan to value ratio (in percentages), and debt to income ratio (in percentages).

Take Panel B, for example, which shows that the 25^{th}, 50^{th} and 75^{th} percentile of credit scores for all loans approved by the other remaining financial institutions in 2002 are 676, 723, and 763. Similarly, the first, second, third quartile combined loan to value for all loans approved by other remaining financial institutions in 2002 are 68, 80, and 85 percent; and the first, second, third quartile debt to income ratio for all loans approved by the other remaining financial institutions in 2002 are 22, 31, and 39 percent. A positive coefficient means a greater than base group, a negative coefficient means a lower than base group. Chase has a coefficient of 0 for the 25^{th} and 50^{th} percentile, which implies that the first and second quartiles of credit score of loans by Chase are the same as the industry average. For the CLTV and DTI ratios, Chase has all negative coefficients, which implies that overall loans approved by Chase in this year have a lower risk than the base group, which explains why Chase has a 7-year cumulative default rate of 1.8%, while the base group has a 7-year cumulative default rate of 2.29%, as shown in Table 5.7.

Citi has a negative coefficient in credit score for all three quartiles; three non-negative coefficients for the CLTV ratio, and a positive DTI ratio, which indicates that loans approved by Citibank had a higher risk than loans approved by the base group. As is shown in Table 5.7, loans originated in 2002 by Citi had a 7-year default rate of 2.15%, lower than the default rate of 2.29% (the base group). This implies these three loan characteristics index can measure most of the loan risk, but cannot explain 100% of the risk, Citi might do better in controlling data quality, management and other aspects of the loan process, which in turn leads to an overall lower default rate than the base group.

From all three perspectives, Bank of America obviously had a lower risk than the base group, which explains why it had a default rate of 0.78, as compared with 2.29% for the base group. The loans by Countrywide have a lower risk than the base group, but with risk higher than the loans by Bank of America, which explains its 7-year default rate of 1.25%, worse than the 0.78% by Bank of America, and greater than 2.29% by the base group.

Table 5.11: Quantile Regression for Three Loan Characteristics across Banks

	Creditscore			CLTV			DTI		
banks	credit(.25)	credit(.5)	credit(.75)	cltv(.25)	cltv(.5)	cltv(.75)	dti(.25)	dti(.5)	dti(.75)
diff of quantile	Panel A: loans originated in 1999(N=934,407)								
Chase	-3	1	4	-5	-2	-9	-2	0	0
Citi	-2	-3	0	3	0	1	2	1	1
Bank of America	3	2	1	-6	-2	-2	3	3	2
Countrywide	0	7	6	-10	-5	-9	-2	-1	-1
all others	675	717	752	70	80	89	23	31	38
diff of quantile	Panel B: loans originated in 2002(N=1,692,187)								
Chase	0	0	1	-9	-7	-3	-2	-2	-1
Citi	-6	-7	-4	1	0	1	3	2	2
Bank of America	21	15	7	-8	-5	-5	1	0	0
Countrywide	12	13	7	-3	-5	-5	0	0	0
Wells Fargo	1	-2	-4	-2	-1	1	0	0	0
all others	676	723	763	68	80	85	22	31	39
diff of quantile	Panel C: loans originated in 2006(N=1,071,788)								
Chase	4	0	0	-2	-3	0	-2	-1	-2
Citi	-20	-18	-10	-12	-6	0	-2	1	0
Bank of America	25	20	10	-7	-4	2	1	1	0
Countrywide	-20	-24	-13	1	-1	0	2	2	1
Wells Fargo	0	-2	0	-2	0	5	2	2	1
all others	685	733	773	62	79	80	22	32	41

Source: Freddie Mac Single Family Loan-Level Dataset all significant at 1% level

In Panel C, it is shown that out of the five big banks, Citi has the lowest CLTV ratio, Bank of America has the highest credit score, and Chase has the lowest DTI ratio, which implies that different banks use different loan risk measurements and have different strategies in controlling loan risk. Loans by Countrywide have the lowest credit score, highest CLTV ratio and highest DTI ratio out of five big banks. With 8 out of 9 characteristic measurements worse than the industry average, it is not surprising that Countrywide had the highest default rate out of all banks. Wells Fargo had an overall higher risk than the industry average loans; 3 zeros show the same as the base group, and 5 measurements are worse than the average. Only 1 measurement, a coefficient of -2 for the 25^{th} percentile CLTV ratio, is better than the base group. However, it has a 7-year default rate of 9.4%, which is lower than the 9.66% for average loans, as shown in Panel C in Table 5.7.

Table 5.12 shows how loan characteristics changed over the years. The five big

Table 5.12: Quantile Regression for Three Loan Characteristics over Years

	creditscore			CLTV			DTI		
banks	credit(.25)	credit(.5)	credit(.75)	cltv(.25)	cltv(.5)	cltv(.75)	dti(.25)	dti(.5)	dti(.75)
diff of quantile	Panel A: loans originated by Chase(N=772,360)								
2000	13	9	7	4	2	9	1	1	1
2001	9	5	5	3	2	10	0	-1	1
2002	4	5	8	-6	-5	2	-1	-2	0
2003	12	9	8	-4	-3	0	-2	-2	0
2004	3	2	5	1	0	0	-1	-1	1
2005	9	5	9	-1	-1	0	0	-1	1
2006	15	13	16	-5	-1	0	-1	0	1
2007	22	18	19	-10	-1	2	-1	-1	1
2008	39	37	28	-6	-3	2	-2	0	2
cons	672	718	756	65	78	80	21	31	38
diff of quantile	Panel B: loans originated by Citi group(N=615,491)								
2000	-5	0	4	2	0	0	1	2	2
2001	-3	0	4	-1	0	0	1	1	1
2002	-3	2	7	-4	0	-4	0	1	2
2003	8	11	11	-5	0	-2	-1	0	1
2004	-7	-5	2	-5	0	0	0	2	2
2005	0	5	12	-8	-1	-5	-1	2	2
2006	-11	-4	8	-24	-7	-10	-5	0	2
2007	-9	1	12	-29	-7	-10	-8	-2	0
2008	21	25	24	-8	-2	-10	-6	-1	1
cons	673	714	752	73	80	90	25	32	39
diff of quantile	Panel C: loans originated by Bank of America(N=537,210)								
2000	-4	-2	-1	3	2	3	-2	-1	0
2001	7	7	7	2	1	-7	-1	-1	0
2002	19	19	17	-4	-3	-7	-3	-3	-1
2003	33	31	25	-10	-8	-7	-5	-4	-3
2004*	a	a	a	a	a	a	a	a	a
2005	33	33	29	-9	-6	-7	-3	-2	0
2006	27	29	27	-12	-4	-7	-3	-1	1
2007	20	27	27	-6	2	3	-6	-3	0
2008	35	35	31	0	2	3	-5	-4	-1
cons	678	719	753	64	78	87	26	34	40
diff of quantile	Panel D: loans originated by Countrywide(N=613,680)								
2000	7	8	10	-6	-1	0	4	3	4
2001	6	6	10	4	0	0	2	2	2
2002	13	12	12	5	0	0	1	1	2
2003	27	16	11	3	0	0	2	1	3
2004	0	-7	1	5	4	0	2	3	4
2005	-6	-14	2	2	3	0	3	4	4
2006	-13	-19	-2	4	3	0	3	4	5
2007	-7	-12	4	6	4	5	2	4	5
2008	12	8	15	8	4	0	1	3	5
cons	675	724	758	60	75	80	21	30	37
diff of quantile	Panel E: loans originated by WellsFargo(N= 3,122,603)								
2001	2	2	3	-3	0	0	-2	-2	0
2002	6	5	6	-8	-1	-4	-4	-3	-1
2003	22	20	15	-13	-5	-10	-8	-6	-3
2004	11	10	9	-10	-1	-4	-4	-2	0
2005	15	15	16	-12	-1	-5	-3	-1	0
2006	11	11	18	-14	-1	-4	-2	0	2
2007	3	4	14	-24	-5	-5	-2	0	2
2008	14	17	22	-8	0	-5	-6	-2	1
cons	671	716	753	74	80	90	26	34	40
diff of quantile	Panel F: loans originated by all others(N=6,730,857)								
2000	-4	0	4	1	0	1	2	2	2
2001	1	5	8	0	0	-4	1	1	2
2002	1	6	11	-2	0	-4	-1	0	1
2003	12	14	15	-5	-3	-9	-1	-1	1
2004	3	6	11	-3	0	-3	0	1	2
2005	8	13	18	-7	-2	-9	0	2	2
2006	7	12	19	-8	-1	-9	-1	2	3
2007	5	10	19	-10	-2	-7	-2	1	3
2008	24	27	27	-6	-2	-7	-3	1	3
cons	675	717	752	70	80	89	23	31	38

Source: Freddie Mac Single Family Loan-Level Dataset

⋆: too few observations

banks are shown from Panel A to E, and the loan characteristics for the industry average is shown in Panel F. In each panel, the base group shows loans originated in 1999. Panel A shows that Chase tightened its lending policies in 2002-2003, credit score coefficients increased, and CLTV and DTI coefficients decreased, all three characteristics indicate tightened lending policies. In 2004, three credit score coefficients decreased from 12, 9, 8 to 3,2, and 5, respectively; three CLTV coefficients increased from -4,-3, 0 to 1, 0, 0 respectively, and three DTI coefficients increased from -2, -2, 0 to -1, -1, and 1 respectively. From 2004 to 2005, the lending policy is mixed; 3 out of 9 coefficients keep constant, and 1 out 9 coefficients changed as the lending policy, and another 5 out of 9 coefficients show a tightening lending policy. After 2005, the coefficients show a clear trend toward a tightening of lending policy. Citi also tightened its lending policies from 2002 to 2003, and then loosened its lending policy from 2003 to 2004; after that Citi tightened its lending policy until 2007. From 2004 to 2007, the CLTV ratio and DTI ratio kept decreasing, and at a very low level.

Bank of America started tightening its lending policies in 2000, and insisted on this policy until 2005. In 2005, both median CLTV and DTI coefficients increased, indicating a loosening lending policy, and after 2005, credit scores continued to decrease (although still very high); both CLTV and DTI keep increasing in 2006, indicating a loosening lending standard.

Countrywide followed a similar trend as Bank of America during the subprime mortgage crisis in loosening its lending policy. Higher risk loans plus an attitude of loosening lending policy led to lower management (less required documentation loans) altogether, which explains the extremely poor performance during the subprime mortgage crisis period.

Wells Fargo followed a similar pattern as Chase and Citi, tightening its lending policy in 2003, and then loosening its lending policy in 2004. After 2005 though, it started tightening its lending policy again.

5.1.3 Missing Data

From 1999-2010, there are 15.85 million loans, and 326,826 observations (2.5% out of total) with credit scores, or loan to value ratio, or debt to income ratio missing. We find that the group with missing data has similar characteristics both over time and across banks. Figure 5.1 shows default comparisons between loans with missing or without missing data over time. It is shown that the overall default rate for both the missing and non-missing groups follows a similar patten. First it decreases from 2000 to 2003, then increases to its peak around 2007, and then decreases again. Loans with missing data have a higher default rate from 2000-2009, the proportion of loans with missing data in 2010 is small, and leads to a lower default rate for loans with missing data than loans with complete information in 2010. Figure 5.2 shows the proportion of loans with missing data over time. The proportion of missing data loans decreased from 2001 to 2005, and then increased from 2005 to 2008, and then decreased again after 2008.

Figure 5.1: Default Comparison Over time

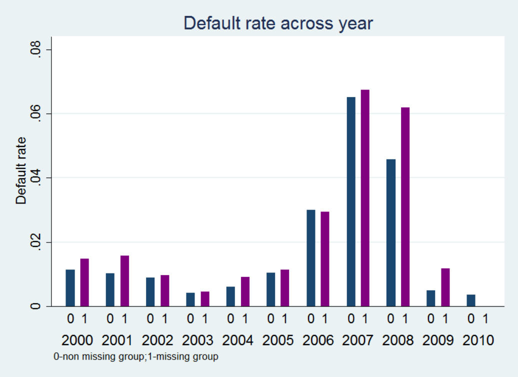


Figure 5.2: Missing Data Proportion over Time

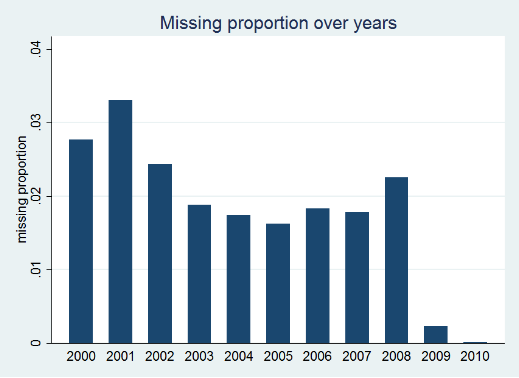


Figure 5.3 shows the default comparisons between loans with missing or without missing data across banks. It shows that most loans with missing data have a slightly higher default rate for most banks, with only two banks different; both Countrywide and Bank of America have a lower default rate for loans with missing data than loans with full information. What can explain this? It signals that both Countrywide and

Figure 5.3: Default Comparison across Banks

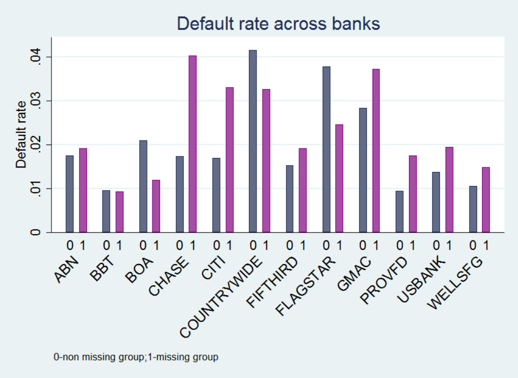
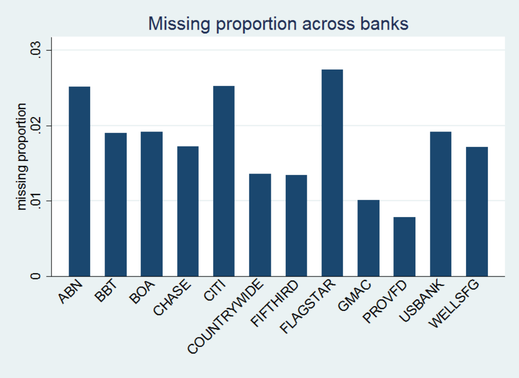


Figure 5.4: Missing Data Proportion for Banks



Bank of America have a lower quality of loan data than other banks. Later we will confirm this based on analysis. Figure 5.4 shows the proportion of loans with missing data across banks. It is shown that Flagstar, Citi, and ABN have about 2.5% of loans with missing data; Countrywide and Bank of America have less than 2% of missing data loans.

5.2 Literature Review and Traditional Theory

5.2.1 Broad Review about Risk Factors

This section presents an overview of the papers published earlier on mortgage default prediction.

Jacobson and Roszbach (2003)discusses how marginal changes in a default risk based acceptance rule would shift the size of the bank's loan portfolio and compares the risk in the sample portfolio with that in an efficiently provided portfolio of equal size. It shows that the size of a small consumer loan does not affect associated default risk.

Agarwal et al. (2012)finds that individual borrower risk characteristics play a significant role in explaining the probability of borrower default, more aggressive mortgage products (hybrid ARMs and no- or low- documentation loans) and increases the probability of borrower default. Extending credit to subprime borrowers in general does not increase the probability of borrower default. *Lawrence et al.* (1992) uses multivariate logit models, investigates the relevance of payment history, loan terms, borrower characteristics, economic conditions, and legal constraints in analyzing loan defaults and delinquencies. It argues that payment history emerges as the overwhelming factor in predicting the likelihood of default.*Kau and Keenan* (1999) argues that house price volatility is quite dramatic, both in terms of severity and the probability of default, and while the volatility of interest rates is more subtle, it cannot be ignored. *Donald et al.* (1996) evaluates the signaling capability of the borrower's s-elected loan to value ratio, and finds the equity proportion of housing capital to be a good indicator of the loan's riskiness. *Lambrecht et al.* (2003) argues that when mortgages are in default, salary and interest are more important than loan to value ratios in influencing the timing decisions of mortgage lenders and borrowers. *Lin et al.* (2011) confirms with *Donald et al.* (1996), and finds that the loan to value ratio and the use status of collateral are significantly positively correlated with the default probability. However, the education degree and the loan amount are significantly negatively correlated with the default probability.

Collateral is one of the important factors. *Chiang et al.* (2002) find that mortgage rates in Hong Kong vary with individual characteristics and a higher mortgage rate is found to be related to either higher collateral (a lower loan-to-value ratio) or slower prepayment. The study suggests that lenders in Hong Kong can observe the risk type of individual borrowers to a certain extent and charge a corresponding mortgage spread.

Neighborhood effect has gotten attention recently. *Chan et al.* (2013) argues that census tract level neighborhood characteristics are important predictors of default behavior based on a database of non-prime mortgages from New York City. They find that default rates increase with the rate of foreclosure notices, and home mortgage defaults are higher in census tracts with larger shares of black residents, regardless of the borrowers own race. *Deng et al.* (2005) also points out that borrowers of similar background tend to cluster together in neighborhoods. The study investigates the impact of spatially correlated unobservable variables on the refinancing, selling and default decisions of mortgage borrowers, and estimates a competing risks hazard model with random effects using a three-stage maximum likelihood estimation approach. It significantly improves the model performance by utilize the space-varying coefficient method.

Race is another significant factor researchers consider about. *Kau et al.* (2012) argues that borrowers in predominantly black neighborhoods pay a significantly higher contract rate than is consistent with evidence of their behavior.

Been et al. (2013)combines data on the performance of mortgage loans with detailed borrower, neighborhood, and property characteristics to examine the factors that determine the outcomes of seriously delinquent loans, and find that the outcomes of delinquent loans are significantly related to: current LTV, FICO scores, especially risky loan characteristics, the servicer of the loan, neighborhood housing price appreciation, and whether the borrower received foreclosure counseling. *Gerardi et al.* (2013) finds no support for the hypothesis that numerical ability impacts mortgage outcomes through the choice of the mortgage contract. The study suggests that individuals with limited numerical ability default on their mortgage due to behavior unrelated to the initial choice of their mortgage.

5.2.2 Paper from Models Perspective

Before going to the model, I will do a brief literature review on default risk models. There is extensive literature on default risk, since *Altman* (1968) proposed the Z-score model, combining traditional ratio "analysis" and discriminant analysis to illustrate the prediction of corporate bankruptcy. Nowadays, since many approaches have been adopted for modelling credit risk, these models improve predictive power. From discriminant analysis to probit, logit regression, from Bivariate probit model to Bivariate logit model, from artificial neural networks to support vector machine, all these models are widely used in predicting default risk. For each of these methods, there is a long list of papers; logistic regression is discussed regarding bankruptcy and credit analysis (*Laitinen and Laitinen* (2001), *Hua et al.* (2007), *Premachandra et al.* (2009)); the bivariate probit model can be seen in *Greene* (1996); the artificial neural networks is discussed in *Desai et al.* (1997), *Tsai and Wu* (2008), and *West* (2000).

However, all of the models mentioned above ignore the information of time to default. In this paper, we introduce survival analysis techniques for modelling mortgage risk. Survival analysis involves the estimation of the distribution of the time it takes for an event to occur to an object, depending on its features. In a medical field, objects often correspond to patients and their features, and are known as explanatory variables and covariates. In the context of loans, the event can be a default, while the loan characteristics are explanatory variables. It attempts to answer questions such as: What is the proportion of loans which will survive past a certain time? How do particular circumstances or characteristics increase or decrease the probability of survival?

The major advantage of survival analysis is that it allows censored data to be incorporated into the model. It focuses the time duration before default happens. The idea of employing survival analysis techniques for constructing credit risk models started with the paper by *Narain* (1992), and then was developed further by *Banasik et al.* (1999), *Stepanova and Thomas* (2002).

Narain (1992) applied the accelerated life exponential model to loan data, and showed that the model estimated the number of failures well. It was pointed out that survival analysis can be applied to any area of credit operations in which there are predictor variables and the time to some event is of interest. *Banasik et al.* (1999)

compared the predictive results of several models, including exponential, Welbull, and Cox's nonparametric and logistic regression, and concluded that survival analysis methods are competitive and superior to the traditional logistic regression.

5.2.3 Traditional Binary Predictive Model

In this section, we begin by briefly introducing multiple discriminant analysis, and then probit and logit analysis. The three models should reveal the same level of statistical significance. Under multiple discriminant analysis, data is assigned to one or more distinct groups. It is appropriate when the groups under examination are discrete and identifiable, each member of the group can be profiled by a set of predictor variables, and the explanatory variables have a normal distribution. Multiple discriminant analysis is theoretically correct only when the grouping populations are normal with identical covariance matrices. However, in the credit default problem, the default variable and many other dummy independent variables are qualitative, which eliminates the possibility of multivariate normality.

5.2.3.1 Probit/Logit Model

The probit model is used to estimate the probability that an observation with particular characteristics will fall into a specific category. It assumes the ordinal nature of the observed variable, and is a popular specification for an ordinal or a binary response model. It is estimated with the standard maximum likelihood method.

Suppose there exists

$$Y_i^* = X_i\beta_i + \epsilon_i fori = 1, 2..., N \tag{5.1}$$

where ϵ is assumed to follow normal distribution. Y can be viewed as an indicator for whether this latent variable is positive:

$$Y_i = \begin{cases} 0 & Y_i^* > 0 \\ 1 & otherwise \end{cases} \tag{5.2}$$

$$L = \prod_{i=1}^{N} Prob(y_i|x_i) = \prod_{i=1}^{N} p(x_i)^{Y_i} * (1 - p(x_i)^{(1-Y_i)} \tag{5.3}$$

$$Pr(Y^* > 0) = Pr(X'\beta + \epsilon > 0) = Pr(\epsilon < X'\beta) = \Phi(X'\beta) \tag{5.4}$$

Pr denotes probability, and Φ is the cumulative distribution function of the standard normal distribution. The parameter β can be estimated by maximizing the joint log-likelihood function.

Equation 5.2 and 5.3 imply the log likelihood

$$lnl = \sum_{i=1}^{N} Y_i ln(\Phi(X'\beta)) + (1 - Y_i) ln(1 - X'\beta) \tag{5.5}$$

Similar to the normal distribution assumption in the probit model, Logit simply assumes a logistic distribution for ϵ.

$$Pr(Y^* > 0) = Pr(X'\beta + \epsilon > 0) = Pr(\epsilon < X'\beta) = logit^{-1}(X'\beta) \tag{5.6}$$

No theoretical and empirical evidence shows that one approach is superior over other approaches when a limited dependent variable model is used. Therefore, a comparison of probit, logit, and multiple discriminant analysis is undertaken to determine whether meaningful different outcomes could be generated in a given empirical problem.

The most basic approach to understanding the classifier results is to consider the number of the predicted default (non-default) and compare this with the actual number of default and non-default. The classifier boundary between default and non-default is determined based on a cut-off value. Table 5.13 shows a contingency table that indicates how the predicted results compare with the actual results.

Table 5.13: Contingency Table

Prediction	Actual(Default)	Actual(Non-Default)
Default	TP	FP
Non-Default	FN	TN
SENS	TP/(TP+FN)	
SPEC		TN/(FP+TN)

In the binary classification, the outcomes are labeled either as positive (default) or negative (non-default). There are four possible outcomes: if both the predicting and actual result is positive, then it is called a true positive (TP); if both the predicting and actual is negative, then it is true negative; if the predicting is positive while the actual is false, then it is false positive (FP); if the predicting is negative while actual is positive, then it is called false negative (FN). FN is a type I error, and FP is a type II error.

In signal detection theory, a receiver operating characteristic (ROC) curve is a graphical plot which illustrates the performance of a binary classifier system as its discrimination threshold varies. ROC analysis provides tools to select possibly optimal models and to discard suboptimal ones. ROC is defined as the the true positive rate (TPR)vs false positive rate(FPR), since TPR is also called sensitivity, and FPR is known as fall-out and can be calculated as one minus specificity. The ROC curve is the sensitivity as a function of fall-out.

The ROC curve depicts relative trade-offs between true positive (benefits) and false positive (costs). For the default prediction model, the relationship is shown in the figure. The 1 unit along x axis represents total actual non-default observation, and the 1 unit along y axis represents total actual default observation. The ROC curve shows that the percentage of defaults can be avoided by setting a cut-off a specific value. ROC provide a pure index of accuracy by demonstrating the limits of a test's ability of discriminate between two groups. The example shows how TP, FP, TN, and FN can be calculated for a specific cut-off value. In the figure, the cut-off is a certain percentile of the model's scores. TP and TN always vary from 0% to 100%, which implies the fact that the baseline default rate in a sample does not change the shape of the ROC.

Figure 5.5: ROC Curve Example-1

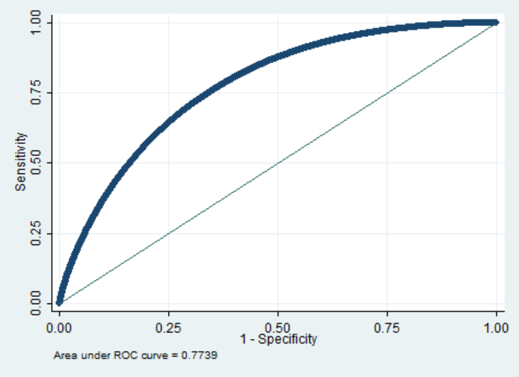
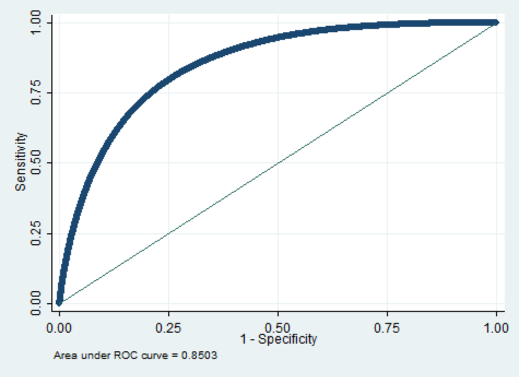


Figure 5.6: ROC Curve Example-2



The bank needs to decide which model/strategy is the best out of all models. The best model should have the highest discrimination power, which can be found by comparing the area under ROC curve with all the models. The bank then can choose an optimal cutoff value to maximize the revenue because the revenue the bank can achieve varies with the choice of different cut-off value for a specific model.

5.2.3.2 Multinomial Logit

When the dependent variables are not 0,1 indicating one event happens or not, but 0, 1,2,3,4, indicating the events can be different types, we use multinomial logit, instead of logit. Under survival analysis, the probability of individual i experiencing j type of events is:

$$\lambda_{i,j} = \frac{e^{X\beta_j}}{\sum_{j=1}^{k} e^{X\beta_k}}$$

X represents a series of explanatory variables, and β represents a vector of coefficients for each of the explanatory variables. β_j represents a set of coefficients for event j. Assume for event 1, it has a set of coefficients of 0, then all the coefficients of explanatory variables would be interpreted with respect to this event 1 baseline event.

Hence, the likelihood function can be written as :

$$L = \Pi_{i=1}^{N}(\lambda_{i1}^{d_{i1}} \times \lambda_{i1}^{d_{i1}} \times ...\lambda_{ik}^{d_{ik}})$$

The Multinomial Logit model for competing risk is the same as before, since duration is not reflected in the likelihood function, The duration is included as one of the explanatory variables.

CHAPTER VI

The Theory of Survival Analysis

Before we start with the discussion of survival analysis, there are several concepts we should know.

Survival analysis concerns analyzing the time to the occurrence of an event. Let T be a nonnegative random variable denoting the time to a failure event. Rather than referring to Ts probability density function, f(t) or F(t), survival analysts talk about Ts survivor function, S(t), or its hazard function, h(t). Although all forms describe the same probability distribution for T, it is more convenient to think of S(t) and h(t).

The survivor function is simply the reverse cumulative distribution function of T:

$$S(t) = 1 - F(t) = Pr(T > t)$$

The survivor function reports the probability that there is no failure event prior to t; in other words, it reports the probability of surviving beyond time t. As t=0, the function is equal to 1 and decreases toward zero as t goes to infinity. The density function, f(t), can be obtained as easily from S(t) as it can from F(t):

$$f(t) = \frac{dF(t)}{dt} = \frac{d}{dt}[1 - S(t)] = -S'(t)$$

The hazard function, h(t), is also known as the conditional failure rate, the intensity function or the instantaneous rate of failure. It is the probability that the failure event occurs in a given interval, conditional upon the subject having survived to the beginning of that interval, divided by the width of that interval:

$$h(t) = \lim_{dt \to 0} \frac{\Pr(t \leq T < t + dt | T \geq t)}{dt} = \lim_{dt \to 0} \frac{\Pr(t \leq T < t + dt)}{dt \cdot S(t)} = \frac{f(t)}{S(t)} = -\frac{S'(t)}{S(t)}$$

The hazard rate can have different shapes, either increase, decrease, remain constant, or even take a more complicated curved shape. The probability of survival past a certain time connects with the amount of risk that has been accumulated up to that time, while the hazard rate measures the rate at which risk is accumulated.

Assume one of the four functions (S,F,f,h)that describes the probability distribution of failure times is given, then the other three function are determined. We may also derive hazard function from the cumulative distribution function:

$$H(t) = \int_0^t h(u) du$$

The cumulative hazard function measures the total amount of risk that has been accumulated up to time t. By a few steps transforming, we can see the relationship between the cumulative hazard function and the survival function.

$$H(t) = \int_0^t \frac{f(u)}{S(u)} du = -\int_0^t \frac{S'(u)}{S(u)} du = -Ln(S(t))$$

6.1 Parametric Models

6.1.1 Maximum Likelihood

Out of all observations, we can assume two completely independent time processes. The first is the actual survival time before the event of interest, the other is the length of time until a subject is lost to follow up. Two variables are used to characterize a subject's actual observed time, T, and a censoring indicator, C. With the covariate variables, X, then for each observation, there are three variables (t, c, x).

First, we need to create the specific likelihood function that yields a quantity similar to the probability of the observed data, then maximize the likelihood function by choosing the appropriate β. We consider two triplets (t, 0, x) and (t, 1, x) separately. In the first case, the survival time is t, and in the second case, the subject is lost to follow, and the survival time is at least t. Hence, for the first case, it is given by the value of density function $f(t, \beta, x)$, and in the second case, it is given by the survivorship function S(t, β, x). Assuming all of the observations are independent, the likelihood function is obtained by multiplying the respective contributions of all the observed triplet, f(t,β, x) for noncensored observations, and S(t, β, x) for censored observations.

In general, the contribution of each triplet to the likelihood function can be written as

$$[f(t, \beta, x)]^c * [S(t, \beta, x)]^{1-c}$$

For a sample of n independent observations as $(t_i, c_i, x_i), i = 1, 2, ...n$, the likelihood function of the entire sample is

$$l(\beta) = \prod_{i=1}^{n} [f(t_i, \beta, x_i)]_i^c * [S(t_i, \beta, x_i)]^{1-c_i}$$

Since the log function is monotone, the maximum likelihood function can be transformed into :

$$l(\beta) = \sum_{i=1}^{n} [c_i Ln[f(t_i, \beta, x_i)] + (1 - c_i) Ln[S(t_i, \beta, x_i)]$$

Now the problem comes to the probability density and survival function form.

As mentioned, parametric models not only differs in terms of the assumptions about density distribution, but also in terms of specification and interpretation; they differ in terms of whether they are proportional hazard (PH) models or accelerated failure models (AFT).

6.1.2 Accelerated Failure Method

Survival analysis concerns analyzing the time to the occurrence of an event. Take an example of mortgage loans: if loans survive x time periods, and with independent variable x, then the simple way to analyze the data is ordinary least-squares linear regression. Since for most survival analysis, if we scatter plot the survival time vs characteristics x, the distribution of survival time appears to the right. The simplest statistical distribution with this characteristic is the exponential distribution.

Suppose there is only one explaining variable, x, then this model can be expressed

$$Ln(t) = \beta x + ln(\varepsilon)$$

$$\varepsilon = e^{-\beta x} t$$

We may also find some sources report the result of:

$$Ln(t) = -\beta x + ln(\varepsilon)$$

While the sign change in β is not important, it would be natural to follow specification without a negative sign, so that a positive coefficient in β increases the value of the log value of time to failure.

With this assumption, we can express the survivor and probability density function and solve the estimated parameter: β and other hazard shaped parameter.

6.1.3 Proportional Hazard Method

The Proportional hazard method assumes that the hazard function is a product of two parts:

$$h(t, X) = h_0(t)e^{X_i\beta} = h_0(T)\lambda$$

$h_0(t)$ is the baseline hazard and depends on t not on X, $\lambda = e^{X_i\beta}$ is a function of characteristics which scales the baseline hazard function up or down. The baseline hazard rate indicates the pattern of time dependency that is assumed to be common to all units. The proportional property indicates that the absolute difference in X implies proportional differences in the hazard rate at each t. When any t,the ratio of hazard rates for two groups i and j with vectors of characteristics X_i and X_j is:

$$\frac{h(\hat{t}, X_i)}{h(\hat{t}, X_j)} = e^{(X_i - X_j)\beta}$$

The proportional difference in the hazard rates of these two groups is fixed across time; in other words, the proportional models assume that the combination of characteristics have a fixed effect across time. It is not a realistic assumption. While it might be true in some cases, most of the time it is not, especially with regard to the analysis of loans originated by different banks, which kept changing strategies about their loan characteristics. It will not be surprising then that this assumption does not hold in such a situation.

In the exponential model, probability density function is assumed to follow exponential distribution.

$$f(t) = \lambda e^{-\lambda t}, \quad t > 0$$

$$F(t) = \int_0^t \lambda e^{-\lambda\tau}\, d\tau = 1 - e^{-\lambda t},$$

$$h(t, X) = \frac{f(t)}{1 - F(t)} = \lambda\_i = e^{X_i \beta}$$

which means that the hazard rate does not change with t, since it is constant over time. Since the density function T has an exponential distribution with mean $1/\lambda$, which can be understood as:

$$E(t_i) = 1/\lambda = \frac{1}{e^{X_i \beta}} = e^{-X_i \beta}$$

If the probability density function is assumed to follow a Weibull distribution, then

$$f(t) = \lambda p (\lambda t)^{p-1} e^{(-\lambda t) p}, \quad t > 0$$

The Weibull model is one with two parameters, λ is the location parameter and p determines whether the hazard rate is increasing, decreasing, or constant over time. When $p > 1$, the hazard rate is monotonically increasing with time; when $p < 1$, the hazard rate is monotonically decreasing with time; if $p = 1$, then the hazard rate is constant as an exponential model.

$$h(t, X) = \lambda p (\lambda t)^{p-1}$$

$$\lambda_i = e^{X_i \beta}$$

With the survival function, the cumulative density function can be solved, and then the probability density function can be estimated. Based on the probability density function and the survival function, the likelihood function can be calculated. Then by maximizing the likelihood function, the coefficient of explanatory variables can be estimated.

All the parametric models we have mentioned can be specified with AFT assumption. However, only the exponential and Weibulls models can be specified in

PH methods. The five parametric models and the specification method are listed in Table 6.1.

Table 6.1: Combination of Model and Specifications

	Proportional Hazard	Accelated Failure Time
Exponential	yes	yes
Webull	yes	yes
Loglogistic	no	yes
LogNormal	no	yes
Log Gamma	no	yes

Table 6.2 shows the comparison of both the hazard function and the expected survival time for the exponential model and the Weibull model with PH and AFT specifications.

Table 6.2: Comparison between PH and AFT for Two Parametric Models

		h(t,X)	E(T)
Exponential	PH	$e^{X_i\beta}$	$e^{-X_i\beta}$
	AFT	$e^{-X_i\beta}$	$e^{X_i\beta}$
Weibull	PH	$e^{X_i\beta}p(e^{X_i\beta}t)^{(p-1)}$	$(\frac{1}{\lambda})^{\frac{1}{p}}\Gamma(1+\frac{1}{p})$
	AFT	$e^{-\frac{X_i\beta}{\sigma}}$	$\frac{\Gamma(1+\frac{1}{p})}{\lambda}$

In Table 6.2, λ is the hazard rate, $exp(X_i\beta)$, which shows that for the exponential model, β from the PH model has a positive effect on hazard rate, β from the AFT model has a negative effect on the hazard; β from the PH model has a negative effect on survival duration, β from the AFT model has a positive effect on the duration. The β under the Weibull model with PH and AFT assumption has a similar effect using the exponential model.

6.2 Non-Parametric Method

Nonparametric analysis makes no assumptions about the function form of the survivor function, letting the dataset speak for itself. The effects of covariates are not modeled either. If the covariates are category data, then the survival experience can be compared at a qualitative level.

6.2.1 The Kaplan-Meier Estimator

The Kaplan and Meier estimator of the survival function is also called the product limit estimator. This estimator incorporates information from all of the observations available, both uncensored and censored. The Kaplan-Meier Estimator survival function is analogous to a toddle who takes several steps to walk from a chair to a table. The second step can only be taken if the first step is successful, and the third step can be taken only if both the first and second steps are successful. In an analysis of survival time, we estimate the conditional probabilities of "successful steps" and then multiply them together to estimate the overall survivor function.

To illustrate these ideas in the context of survival analysis, we describe estimation of the survivor function in detail using data for 5 loans example, as shown in Table 6.4.

Table 6.3: Survival Time and Default Status for Five Loans

id	t	failed	censored
1	2	1	0
2	5	0	1
3	6	1	0
4	8	1	0
5	12	0	1

The "steps"are intervals defined by a rank ordering of the survival times. Subject 1's survival time of 2 months is the shortest and is used to define the interval $I_0 = [0, 2)$. The second rank-ordered time is subject 2's censored survival time of 5 months. The survival time, in conjunction with the ordered survival time of subject 1, defines interval $I_1 = [2, 5)$, the next ordered time is $I_2 = [5, 6)$. Similarly, $I_3 = [6, 8)$, $I_2 = [8, 12)$, and the last interval is defined as $I_5 = [12, +\inf)$.

All subjects are alive at time $t = 0$ and remained so until subject 1 defaults at 2 months. Thus, the estimate of the probability of surviving through interval I_0 is 1.0, and the estimated of survivor function is

$$\hat{S}(t) = 1.0$$

at each t in I_0. Just before time 2 months, five subjects were alive, and at 2 months, one loan defaulted. In order to describe the value of the estimator at 2 months, consider a small interval beginning just before 2 months and ending at 2 months. This can be written as $(3 - \delta, 3]$. The estimated conditional probability of default (dying) in this small interval is 1/5 and the probability of surviving through this interval is $1 - 1/5 = 0.8$. At any specified time point, the number of subjects alive is called the number at risk. At time 2 months, this number is denoted as n_1, the 1 referring to the fact that 2 months is the first observed time. The number of default at 2 months was 1, but with a larger sample, more than one could have been observed. We denote the number of default as d_1. In this more general notation, the estimated probability of surviving is $(n_1 - d_1)/n_1$. The probability that a subject survives to 2 months is estimated as

$$\hat{S}(2) = 1.0 * (4/5) = 0.8$$

The number at risk at the next observed time, 5 months, is $n_2 = 4$, and the number of defaults (death) is zero, since subject 2 was lost to follow due to other reasons. The estimated conditional probability of survival through a small interval is $(4 - 0)/4 = 1.0$. Hence, the estimated survivor function is obtained by successive multiplication of the estimated conditional probabilities and is:

$$\hat{S}(5) = 1.0 * (4/5) * 1 = 0.8 * 1 = 0.8$$

The survivor estimator ar any point in time is obtained through multiplying a sequence of conditional survival probability estimators. Each conditional probability

Table 6.4: KM Survival Function Estimator Example

id	t	failed	censored	$\hat{s}(t)$
1	2	1	0	1.0*(4/5)=0.8
2	5	0	1	1.0*(4/5)*1=0.8
3	6	1	0	1.0*(4/5)*1*(2/3)=0.536
4	8	1	0	1.0*(4/5)*1*(2/3)*(1/2)=0.268
5	12	0	1	1.0*(4/5)*1*(2/3)*(1/2)*1=0.268

estimator is obtained from the observed number at risk and the actual observed default, which is equal to $(n-d)/n$.

For a dataset with observed failure times, $t_1, ...t_k$ where k is the number of distinct failure times observed in the data. The Kaplan-Meier estimator of survivor function at time t is obtained from the equation:

$$\hat{S}(t) = \prod_{j|t_j \leq t} \frac{n_i - d_j}{n_j}.$$

where n_j is the number of subjects at risk at time t_j and d_j is the number of failures at time t_j. The product is over all observed failure times less than or equal to t.

The standard error reported for the Kaplan-Meier estimate is that given by Greenwood's formula:

$$V\hat{a}rS(t) = \hat{S}^2(t) \sum_{j|t_j \leq t} \frac{d_j}{n_j(n_j - d_j)}$$

The method shown above to derive the estimator is the "traditional"approach, and may be found in most texts on survival analysis published prior to 1990. However, these standard errors are not used for confidence intervals. Instead, the asymptotic of $ln(-ln\hat{S}(t))$,

$$\hat{\sigma}^2(t) = \frac{\sum_{j|t_j \leq t} \frac{d_j}{n_j(n_j-d_j)}}{[ln(\hat{S}(t))]^2}$$

is used. It was originally proposed by *Hall and Wellner* (1980), and later was discussed in detail by many others.

6.2.2 The Nelson-Aalen Estimator

The cumulative hazard function is defined as

$$H(t) = \int_{0}^{t} h(u)du$$

There is a theoretical relationship between H(t) and S(t),

$$H(t) = -lnS(t)$$

where for S(t) we could use the Kaplan-Meier estimator. *Nelson* (1972) and *Aalen* (1978) proposed another nonparametric method for estimating H(t), because it has better small-sample properties.

$$\hat{H}(t) = \sum_{j|t_j \leq t} \frac{d_j}{n_j}$$

where n_j is the number at risk at time t_j, d_j is the number of failures at time t_j, and the sum is over all distinct failure times less than or equal to t. This gives the same data sample in the last section, so we can calculate the number of failures per subject at each observed time, $e_j = d_j/n_j$, and then sum these to form $\hat{H}(t)$.

Table 6.5: Nelson-Aalen Cumulative Hazard Estimator Example

id	t	failed	censored	e_j	$\hat{H}(t)$
1	2	1	0	1/5	0.2
2	5	0	1	0	0.2
3	6	1	0	1/3	0.53
4	8	1	0	1/2	1.03
5	12	0	1	0	1.03

The standard errors reported are based on the variance calculation (*Aalen* (1978)),

$$\hat{Var}[\hat{H}(t)] = \sum_{j|t_j \leq t} \frac{d_j}{n_j^2}$$

and the confidence intervals reported are $\hat{H}(t) exp \pm z_{\alpha/2}\hat{\phi}(t)$,

$$\hat{\phi^2} = \frac{\hat{Var}[\hat{H}(t)]}{[\hat{H}(t)]^2}$$

Table 6.6: Kaplan-Meier and Nelson-Aalen Estimator Comparison

id	t	failed	censored	S_{KM}	S_{NA}	H_{NA}	H_{KM}
1	2	1	0	0.8	0.819	0.2	0.223
2	5	0	1	0.8	0.819	0.2	0.223
3	6	1	0	0.536	0.589	0.53	0.623
4	8	1	0	0.268	0.357	1.03	1.317
5	12	0	1	0.268	0.357	1.03	1.317

By the theoretical relationship between H(t) and S(t),

$$S(t) = e^{-H(t)}$$

we can calculate the survival function from the Nealson-Aalen cumulative hazard function, or calculate the cumulative hazard function from the Kaplan-Meier survival function. Table 6.6 shows the survival function and cumulative hazard function from both methods. We can see that the KM survivor function is less than the NA transforming survivor function, and NA cumulative hazard function is less than the KM transforming cumulative hazard function. Actually, it is not hard to prove that this is always this case for any survival analysis.

6.2.3 Testing the Equality of Survival Functions

The estimated survivor function for one group might lie completely above the other group. In general, the pattern of one survivor function lying above another means the group defined by the upper curve has a more favorable survival experience, since the upper group lives longer. In other words, at any point in time the proportion of subjects estimated to be alive is greater for the upper curve group than for the lower curve group.

The statistical question we need to ask is whether the observed difference between different groups is significant. Many statistical tests have been proposed to answer this question.

To make things easier, we start with two groups, group 1 and 2. Assuming that there are k distinct failure times, and at failure time $_j$, there are d_j fails, the total number of subjects that are at risk is n_j, so $n_j - d_j$ survive.

Table 6.7: Test of Equality of Survivor Function in Two Groups

Group	Failure at t_j	survive at t_j	at risk
1	d_{1j}	$n_{1j} - d_{1j}$	n_{1j}
2	d_{2j}	$n_{2j} - d_{2j}$	n_{2j}
total	d_j	$n_j - d_j$	n_j

The contribution to the test statistic depends on which of the various tests is used, but each may be expressed in the form of a ratio of weighted sums over the observed survival times. These tests may be defined in general as:

$$Q = \frac{[\sum_{j=1}^{k} w_j(d_{1j} - \hat{e}_{1j})]^2}{\sum_{j=1}^{k} w_j^2 \hat{v}_{1j}}$$

$\hat{e_j}$ is the estimator default number for different groups, assuming that the survivor function is the same with the overall survivor function. For group 1, the estimator is

$$\hat{e}_{1j} = n_{1j} * \frac{d_j}{n_j}$$

and

$$\hat{V_{1j}} = \frac{n_{1j} n_{2j} d_j (n_j - d_j)}{n_j^2 (n_j - 1)}$$

This is the variance estimator of d_{1j} on the hypergeometric distribution defined by most software packages.

Under the null hypothesis that the two survivor functions are the same, and assuming that the censoring experience is independent of the groups, and that the total number of observed events and the sum of the expected number of events is large, then the significance level for Q may be obtained using the chi-square distribution with one degree of freedom [$p = Pr(\chi^2(1) \geq Q)$].

For the log rank test, which can be seen as an extension of the familiar Mantel-Haenszel test (*Mantel and Haenszel* (1959)), Wt_j is assumed to be 1. The Wilcoxson test, discussed in *Gehan* (1965) and *Breslow* (1970), is constructed in the same way, except that $Wt_j = n_j$, which places more weight at earlier failure times because there are more subjects at risk at earlier times. The Tarone-Ware test, based on the work of *Tarone and Ware* (1977) is also identical to two other tests, except that $Wt_j = \sqrt{n_j}$. The Peto-Peto-Prentice test, based on *Peto and Peto* (1972) and *Prentice* (1978) , uses $Wt_j = \hat{S}_j$.

Table 6.8: Test of Equality of Survivor Function

test	$w(t_j)$
Log rank test	1
Wilcoxon test	n_j
Tarone-Ware test	$\sqrt{n_j}$
Peto-Peto-Prentice test	$\hat{S}(t_j)$

6.3 Semi-Parametric Model:Cox Proportional Model

The Cox Proportional model is a special case for the proportional hazard (PH) model, assuming that the hazard rate for the jth subject in the data is

$$h(t|x_j) = h_0(t) exp(x_j \beta_x)$$

where the baseline hazard, h_0, is not specified in any particular function and not estimated. The model makes no assumption about the shape of the hazard over time' it could be increasing, decreasing, constant, increasing and then decreasing, decreasing then increasing, or any other shape. However, whatever the shape is, it is the same for all subject. One subject's hazard is always a multiplicative replica of another's. The Cox model has no intercept because any intercept is absorbed into the baseline hazard h_0.

$$h(t|x_j) = h_0exp(\beta_0 + x_j\beta_x) = [h_0(t)exp(\beta_0)]exp(x_j\beta_x)$$

Compare the hazard rate for two subjects, j and m, then

$$\frac{h(t|x_j)}{h(t|x_m)} = \frac{exp(x_i\beta_x)}{exp(x_m\beta_x)}$$

Assume the covariates x_j and covariates x_m do not change over time, then the above ratio is constant.

If we cannot make reasonable assumptions about the hazard function, then leaving it unspecified is the best choice. Compared with the parametric method, making assumptions about the shape of the hazard, such as $h_0(t) = a$ or $h_0(t) = apt^{p-1}$, the Cox proportional model gives a considerable advantage when we are uncertain about the hazard function. In the parametric model, if we make a wrong assumption about the hazrad function, then the results could be misleading; however, we won't make such mistakes in the Cox proportional model because we simply make no assumption about the hazard function.

Although the Cox model makes no assumption about the baseline hazard, estimates of function can be obtained after β_x is estimated. When $x = 0$, the relative hazard is 1, then the baseline hazard function can be estimated. We may also obtain estimates of the baseline survivor function $S_0(t)$ and the baseline cumulative hazard function $H_0(t)$ corresponding to $h_0(t)$.

6.3.1 Maximum of Partial Likelihood

Cox (1972) proposed using an expression of "partial likelihood function" that depends only on the parameter of interest. He speculated that the resulting parameter estimators from the partial likelihood function would have the same distribution properties as the full maximum likelihood estimators. Rigorous mathematical proofs have been done and later work simplified the earlier work. The essential idea is simple: the partial likelihood is given by the expression:

$$L(\beta) = \prod_{j=1}^{n} [\frac{e^{x_j\beta}}{\sum_{i\in R(t_j)} e^{x_i\beta}}]^{c_j}$$

where the summation in the denominator is over all subjects in the risk set time t_i, denoted by $R(t_i)$. Recall that the risk set consists of all subjects with survival or censored times greater than or equal to the specified time.

The expression assumes that there are no tied times, and it is often modified to exclude terms when $c_j = 0$, yielding

$$L(\beta) = \prod_{j=1}^{k} \frac{e^{x_j\beta}}{\sum_{i\in R(t_j)} e^{x_i\beta}}$$

where the product is over the k distinct ordered survival times and x_j denotes the values of the covariate for the subject with ordered survival time t_j. The log partial likelihood function is

$$LL = \sum_{j=1}^{k} [x_j\beta - ln[\sum_{i\in R(t_j)} e^{x_i\beta}]]$$

differentiating the right hand side of the log likelihood function with respect to β, setting the derivative equal to zero and solving for the unknown parameter.

We will use an example in *Cleves* (2008), shown in Table 6.9, to demonstrate how the Cox model estimates the covariate coefficient.

The data include 4 failure subjects, and is ordered according to their survival

Table 6.9: Cox Proportional Example Data - without Tie of Failures

id	t	x
1	2	4
2	3	1
3	6	3
4	12	2

time. The first subject fails at time 2, and the second subject fails at time 3, the third subject fails at time 6, and the last subject fails at time 12. There are four distinct risk pools:

at time 2: Subjects are at risk:1,2,3,4, observed to fail:1;

at time 3: Subjects are at risk:2,3,4, observed to fail:2;

at time 6: Subjects are at risk:3,4, observed to fail:3;

at time 12: Subjects are at risk:4, observed to fail:4;

At each of the failure times, we calculate the conditional probability of failure for the subject who actually is observed to fail. The likelihood function can be written as:

$$L(\beta) = P_1 P_2 P_3 P_4$$

where each P_i, $i = 1, ...4$ represents a conditional probability for each failure time. Given the fact that one of the subjects must fail, at the last time period($t = 12$), there is only one subject at risk, and the probability of observing the failure subject 4 is 1, $p_4 = 1$. At $t = 6$, there are two subjects at risk, and the probability of observing subject 3 is:

$$P_3 = \frac{h(6|x_3)}{h(6|x_3) + h(6|x_4)} = \frac{exp(x_3\beta)}{exp(x_3\beta) + exp(x_4\beta)}$$

In the partial likelihood function, the baseline hazard is the same and canceled out. This is fundamental to the Cox proportional model; the order of the failure time matters, not the actual times themselves. Similarly, the likelihood of P_1 and P_2.

$$P_2 = \frac{exp(x_2\beta)}{exp(x_2\beta) + exp(x_3\beta) + exp(x_4\beta)}$$

$$P_1 = \frac{exp(x_1\beta)}{exp(x_1\beta) + exp(x_2\beta) + exp(x_3\beta) + exp(x_4\beta)}$$

Substitute P_1,...P_4 in the log likelihood function, then the maximum log likelihood we can estimated β.

In the example given before, there are no two subjects failing at the same time, so this assumption makes the estimate easy. When there is a tie of failures, it becomes more complicated.

$$Pr(jfails|risksetR_j) = \frac{exp(x_j\beta_x)}{\sum_{i\in R_j exp(x_i\beta_x)}}$$

Introducing the notation $r_j = exp(x_j\beta)$ can make the above formula more compact as:$Pr = r_j/\sum_{i\in R_j r_i}$

Now assume two subjects fail at the same time, as shown in Table 6.10.

Table 6.10: Cox Proportional Example Data - with Tie of Failures

id	t	x
1	2	4
2	3	1
3	3	3
4	12	2

In the new data set, we can see that: at time 2: Subjects are at risk:1,2,3,4, observed to fail:1;

at time 3: Subjects are at risk:2,3,4, observed to fail:2,3;

at time 12: Subjects are at risk:4, observed to fail:4.

How do we calculate the probability that both subject 2 and 3 fail given that these two subjects fail at the same time? We know subject 2 and 3 did not really fail at the same time, we were limited to how precisely we could measure the data. There are two possibilities; one is subject 2 fails and then subject 3 fails after that. The other is that subject 3 fails first then subject 2 fails after. Define P_{23} as the probability that subject 2 fails and then subject 3 fails, andP_{32} as the probability that subject 3 fails and then subject 2.

$$P_{23} = \frac{r_2}{r_2 + r_3 + r_4} \frac{r_3}{r_3 + r_4}$$

$$P_{32} = \frac{r_3}{r_2 + r_3 + r_4} \frac{r_2}{r_2 + r_4}$$

If we know the exact ordering then choosing either P_{23} or P_{32} to substitute $P_2 P_3$ would represent the two middle failure times. However, since we do not know the exact order, we can instead take the probability $P_{23} + P_{32}$ to substitute the two middle failure subjects. This method of calculating the conditional probability of tied failure events is called the marginal calculation, the exact-marginal calculation.

Another way is to assume that the failures really occur at the same time and treat this as a multinominal problem. Given that two subjects are to occur at the same time amon subjects 2, 3,4, the possibilities are: 2 and 3 fail, or 2 and 4 fail, or 3 and 4 fail. The conditional probability that 2 and 3 are observed from this set of possibilities is

$$P_{23} = \frac{r_2 r_3}{r_2 r_3 + r_2 r_4 + r_3 r_4}$$

This method is known as the partial calculation, the exact-partial calculation, or the discrete time calculation, or the conditional logistic calculation.

Both the exact marginal and partial calculations are very computationally intensive so it has become popular to use approximations. *Breslow* (1974) and *Efron* (1977) proposed different approximations of exact marginal method. With the Breslow approximation, the risk pool for the second and subsequent failure events within a set of tied failures is not adjusted for previous failures. The Breslow method uses:

$$P_{23} = \frac{r_2}{r_2 + r_3 + r_4} \frac{r_3}{r_2 + r_3 + r_4}$$

$$P_{32} = \frac{r_3}{r_2 + r_3 + r_4} \frac{r_2}{r_2 + r_3 + r_4}$$

with this approximation, the denominator is the same, the calculation is simplified considerably.

$$P_{23} + P_{32} = 2r_2r_3/(r_2 + r_3 + r_4)^2$$

The Efron approximation adjusts the subsequent risk sets using probability weights. At time 3, after one of the tie failures occurs, the second risk set is either $3, 4$ or $2, 4$, so the approximation uses the average of the two sets, $(r_2 + r_4 + r_3 + r_4)/2 = (r_2 + r_3)/2 + r_4$. Hence

$$P_{23} = \frac{r_2}{r_2 + r_3 + r_4} \frac{r_3}{\frac{1}{2}(r_2 + r_3) + r_4}$$

$$P_{32} = \frac{r_3}{r_2 + r_3 + r_4} \frac{r_2}{\frac{1}{2}(r_2 + r_3) + r_4}$$

$$P_{2*}P_3 = P_{23} + P_{32} = \frac{2r_2r_3}{(r_2 + r_3 + r_4)[\frac{1}{2}(r_2 + r_3) + r_4]}$$

The Efron approximation is more accurate than Breslow's approximation, but it takes longer to calculate.

6.3.2 Cox Model Diagnostics

Just like any ordinary least-squares model, the Cox proportional model needs to check model specification, goodness to fit, outlier and influential points. The specifications help us search for variables to add to the model. If the model is already correctly specified, then the added variables will add little or no explanatory power, so we can test whether variables are “insignificant”.

6.3.2.1 The Link Test

One easy and powerful way to verify whether we include the appropriate explaining variables is the link test, it is true for all regression models, and not unique to the Cox model. The link test verifies that the coefficient on the squared linear predictor is insignificant. With this test, we first estimate β_x from the standard Cox model and then estimate β_1 and β_2 from a second round model

$$LHR = \beta_1(x\hat{\beta})_x + \beta_2(x\hat{\beta}_x)^2$$

Under the assumption that $x\beta_x$ is the correct specification, $\beta_1 = 1$, and $\beta_2 = 0$, we test that $\beta_2 = 0$.

This test can be slightly modified to interact time analysis with the covariates and verify whether the effects of these interacted variables are no different from zero because the proportional hazard model assumes the effect does not change with time. If we have the correct specification, the effect will not change with time.

$$LHR = beta_1(x\hat{\beta})_x + \beta_2(x\hat{\beta}_x)t$$

We can rewrite the model as:

$$LHR = x\beta_{x1} + x\beta_{x2}t$$

and test whether $\beta_{x2} = 0$. The most popular method is to fit one model per covariate, and then test separately. The basis of the specification test is the consideration of models of the form

$$LHR = x_j\beta_x + \beta_2 q_j$$

Under the assumption that $x\beta_x$ is the correct specification, β_2 will be zero.

6.3.2.2 Martingale Residuals

Martingale residuals can be interpreted simply as the difference between the observed number of failures in the data and the number of failures predicted by the model. Martingale residuals help us to determine what is the best function form to use for each covariate. Assuming $h(t|x_i) = h_0(t)exp(f(x_i))$, x_i is the covariate vector for the ith subject and f() is some function. Let M_i be the Martingale residual of the ith observation obtained when no covariates are included in the model. Then Mi is approximately $kf(x_i)$, where k is a constant that depends on the number of censored observations. The approximation shows there is a linear relationship between M_i and $f(x_i)$. Then we can check whether the Martingale residual is linear with the covariates that we should include (from general link test),if not linear with covariate itself, then try the log and other functions of the covariate. We will apply this test in the application section.

6.3.2.3 Schoenfeld Residuals

One way to check the proportional hazard assumption (specification) is by introducing time in the link test. Another way to test is based on the Schoenfeld residual, proposed by *Schoenfeld* (1982), when there are no tied failure times, the simplest form is:

$$r_{uj} = x_{uj} - \frac{\sum_{i\in R_j} x_{ui} exp(x_i\hat{\beta_x})}{\sum_{i\in R_j} exp(x_i\hat{\beta_x})}$$

The Schoenfeld residual, r_{uj} , for covariate $x_u, u = 1, ..p$, and for subjects j is the difference between the covariate value for the failed observation and the weighted average of the covariate values (weighted according to the estimated relative hazard from the Cox model) over all those subjects at risk of failure when subject j failed.

If the assumption does not hold, the coefficient on x_u vary with time,

$$\beta_u(t) = \beta_u + q_j g(t)$$

where g is the function of time, q is the coefficient. With proportional hazard assumption, $q_j = 0$. *Grambsch and Therneau* (1994) provides a method of scaling the Schoenfeld residual to form r^*_{uj}, and

$$E(r^*_{uj} + \beta_u) = \beta_u(t)$$

We can plot a graph of r^*_{uj} versus t_j, or some function of t_j, it should show a zero slope if the assumption is correct.

Another graphical method is proposed by *Hess* (1995) in which they plot an estimate of $\ln[-ln(\hat{S}(t))]$ versus $ln(t)$ for each level of the covariate in question, where $\hat{S}(t)$ is the Kaplan-Meier estimate of the survivor function. Under the proportional hazard assumption, the plotted curves should be parrallel.

6.3.2.4 Cox Snell Residuals

To evaluate the overall fitness, we use Cox-Snell residuals ().

$$CSr_j = \hat{H}_0(t_j) exp(x_j \hat{\beta}_x)$$

where both $\hat{H}_0$ and $\hat{\beta}_x$ are obtained from the Cox model estimate. If the Cox model fits the data, then the true cumulative hazard function conditional on the covariate vector has an exponential distribution with a hazard function equal to 1 for all t, and the cumulative hazard of the Cox-Snell residuals should be a straight line with a slope of 1.

We can also assess the predictive power of the Cox model by computing the Harrell's C concordance statistics or Somers'D statistics. This measures the agreement of the predictions with the observed failure order, defined as the proportion of all usable subjects pairs in which the predictions and outcomes are concordant. In the case of Harrells c, the estimated parameter is based on a scale of 0 to 1, and is expected to be at least 0.5 for a positive predictor of lifetime, such as an inverse hazard ratio. In the case of Somers D, the untransformed parameter is on a scale from -1 to 1, and is expected to be at least 0 for a positive predictor of lifetime. The two measures are closely related: $D = 2(C - 0.5)$.

6.4 Competing Risk Model

Competing risk refers to the chance that instead of default, we will also observe a competing event, for example, prepayment or foreclosure. The competing event, whether prepayment or foreclosure, impedes the occurrence of the event of interest, default. This is different from the right-censoring in the survival data. When subjects are lost to follow-up, they are still considered ar risk of recurrent default, but the researcher is not in a position to record the precise time that it happens. In contrast, prepayment is a permanent condition that prevents future default. While censoring merely obstructs one from observing the default event, a competing event prevents the default event from occurring altogether.

When competing risks are present, cumulative incidence function (CIF) should be focused; in other words, we should focus on the failure function P($T \leq t$ and default), instead of focusing on the survivor function P($T > t$ and default). This is because we don't know what time the event will occur until after it has actually occurred. It makes more sense to ask, "What is the probability of default within n years?" than to ask "What is the probability that nothing happens before n years, and that when something does happens, it will be default and not prepayment?".

Traditionally, for the standard survival analysis, we use 1 minus the Kaplan- Meier estimator to estimate the cumulative incidence failure. Adapting this estimator for competing risks data is not as simple as treating a competing risk event as censored. $1 - \hat{S}_k m$ is biased for two reasons. First, it does not consider the possible correlation between competing risks. Second, the estimator always begins at zero at the beginning and eventually approach 1 if not censored. However, with competing events, $CIF_i(t)$ is strict less than 1, and it will approach the probability of eventual failure from cause i.

Mathematically,

$$CIF_i(t) = \int_0^t h_i(x)S(x)dx = \int_0^t h_i(x)exp[-\sum_{j=1}^{k} H_j(x)]dx \tag{6.1}$$

$S(x)$ is the overall survivor function, and the probability of being failure-free from any cause up to time x. $H_j(x)$ is the cause-specific cumulative hazard for cause j, the sum of all $H_j(x)$ for all different types of failure cause is the overall cumulative hazard. It shows that CIF for cause i is not just a function of the cause-specific hazard function for cause i, but instead is a function of all cause-specific hazards.

6.4.1 Nonparametric Analysis

The nonparametric method makes no assumption about the functional form of the hazard and make no assumption about how hazards differ among groups; it lets the data speak for itself.

The estimate of cumulative incidence is based on 6.1, namely,

$$C\hat{I}F_i(t) = \sum_{j:t_j \le t} \frac{d_{ij}}{n_j} \hat{S}(t_{j-1})$$

where $\hat{S}()$ is just the Kaplan Meier estimate of survival from all failure causes, which can be computed directly by treating all failures types as just"failure". The sum is over all times $t_j \leq t$; where a failure from cause i takes place. d_{ij} is the number of failures from cause i at time t_j, and n_j is the number at risk of failing from any cause at time t_j. The above estimator is from *Marubini and Valsecchi* (2004), and details can be see in *Coviello and Boggess* (2004).

6.4.2 Semiparametric Analysis

One representative of the competing risk scenario is described by *Lunn and McNeil* (1995) and *Cleves* (2000). They both fit regression simultaneously by performing some data duplication and manipulation beforehand. If there are k types of events, then we duplicate dataset k times, and establish a variable of "type", which is denoted from 1 to k for each of the duplicated subjects. We define the fail as 1 if the type variable equals to the event status. We specify "type" as an interaction with covariates, allowing the covariate effect to change for different failure types. Then we fit the two simultaneous Cox regression by completely stratifying on type. By using this technique, we can find a model for both cause-specific hazards by imposing some structure.

Another way is to estimate the hazard rate for different types of events separately, assuming there are 2 types of events, then applying the Cox-proportional model to perform regression on event 1 by treating failures of type 2 as censored, then on event 2 by treating failure of type 1 as censored. We can estimate the hazard for the event of interest, $h_1(t)$ and the hazard for the competing event, $h_2(t)$, from the available data. Then we can combine $h_1(t)$ and $h_2(t)$ to form a total hazard, h(t), so that any event will occur. According to the equation 6.1, we can then calculate cause specific cumulative hazard.

The last method of the competing risk model is proposed by *Fine and Gray*

(1999), in which they specify a model for subdistribution hazard, making it easy to see the effect of the covariates. For the failure of type 1 events, the hazard of the subdistribution is defined as

$$h_1(t) = lim_{\delta \to 0}\{\frac{P(t < T \leq t + \delta \, and \, event \, type 1)|T > t or (T \leq t \, and \, not \, event \, type 1)}{\delta}\}$$

This hazard generates failure events of the interest (default) while keeping the loans that experience competing events "at risk" so that they can be adequately counted as not having any chance of failing. The advantage of the subhazard is that we can readily calculate the CIF from $\bar{H}_1(t)$, the cumulative subhazard.

$$CIF_1 = 1 - exp\{-\bar{H}_1(t)\}$$

This mechanism used to do competing risk regression is similar to the Cox-proportional model. The model is semiparametric in that the baseline hazard $h_{1,0}(t)$ (all the covariates set to zero) is left unspecified, while the effects of the covariates are assumed to be proportional:

$$\bar{h}_1(t|x) = \bar{h}_{1,0}(t)exp(x\beta)$$

For the parametric model, we need to assume the function form, and then do a regression. We may think of fitting a Weibull distribution for subhazard model, then with maximum likelihood function, solve the coefficient and distribution parameter.

CHAPTER VII

Application of Survival Analysis

In this section, we apply survival analysis to all of the loans originated in the state of Florida during 1999-2008 period. Since the loans originated in 2008 only have five years of performance data, we will follow the 5-year performance for all of the loans.

Table 7.1: Loans Originated in Florida from 2000-2008

origyear	Base	Chase	Citi	BoA	CountryWide	Wells Fargo	Total
2000	23,117	5,122	1,975	7,172	1,371	7,237	45,994
2001	47,874	7,726	3,458	8,768	270	21,417	89,513
2002	56,179	4,559	3,777	12,984	320	32,406	110,225
2003	54,412	12,451	4,150	3,183	1,513	46,379	122,088
2004	30,487	22,708	3,061	11	565	27,319	84,151
2005	36,830	17,746	3,143	3,923	5,779	28,331	95,752
2006	32,407	11,153	1,508	3,872	7,732	22,324	78,996
2007	29,776	6,901	1,021	10,772	8,750	11,523	68,743
2008	24,038	7,445	2,227	9,875	3,557	8,432	55,574
Total	335,120	95,811	24,320	60,560	29,857	205,368	751,036

Table 7.1 shows that there are 751,036 loans originated in Florida during 2000-2008, and indicates the number of loans originated by different banks across years. It is shown that the number of loans are not uniformly distributed over time. Overall, banks approved large number of loans during 2002-2003, and a smaller number of loans during 2007-2008. However, for different banks, the situation is different. Chase approved the largest number of loans during 2004-2005, Bank of America approved more loans during 2007-2008, and Countrywide approved even more loans during

2006-2007.

7.1 Nonparametric Survival Function

Table 7.2: Kaplan-Meier Survivor and Nelson-Aalen Cumulative Hazard Function

Time	Beg Total	Fail	Survivor Func	Std Error	[95% Conf. Int.]		N-A Cum.Haz.	std Error	[95% Conf. Int.]	
1	618741	0	1.0000	.	.	.	0.0000	0.0000		
6	599193	331	0.9994	0.0000	0.9994	0.9995	0.0006	0.0000	0.0005	0.0006
11	556943	2050	0.9959	0.0001	0.9957	0.9960	0.0041	0.0001	0.0040	0.0043
16	494688	3248	0.9896	0.0001	0.9894	0.9899	0.0104	0.0001	0.0101	0.0107
21	428914	4763	0.9793	0.0002	0.9789	0.9797	0.0209	0.0002	0.0205	0.0213
26	369850	5606	0.9653	0.0003	0.9648	0.9659	0.0352	0.0003	0.0347	0.0358
31	323407	6495	0.9470	0.0003	0.9464	0.9477	0.0543	0.0004	0.0536	0.0551
36	286285	6618	0.9264	0.0004	0.9255	0.9272	0.0764	0.0005	0.0755	0.0773
41	256152	6894	0.9027	0.0005	0.9017	0.9037	0.1021	0.0006	0.1011	0.1032
46	228888	6913	0.8770	0.0006	0.8758	0.8781	0.1310	0.0007	0.1298	0.1323
51	204848	6312	0.8514	0.0006	0.8502	0.8527	0.1605	0.0008	0.1590	0.1620
56	182150	5877	0.8256	0.0007	0.8242	0.8269	0.1912	0.0009	0.1896	0.1929

Source: Freddie Mac Single Family Loan-Level Dataset

Table 7.2 shows the nonparametric Kaplan Meier survivor function.

$$\hat{S}(t) = \prod_{j|t_j \leq t} \frac{n_i - d_j}{n_j}.$$

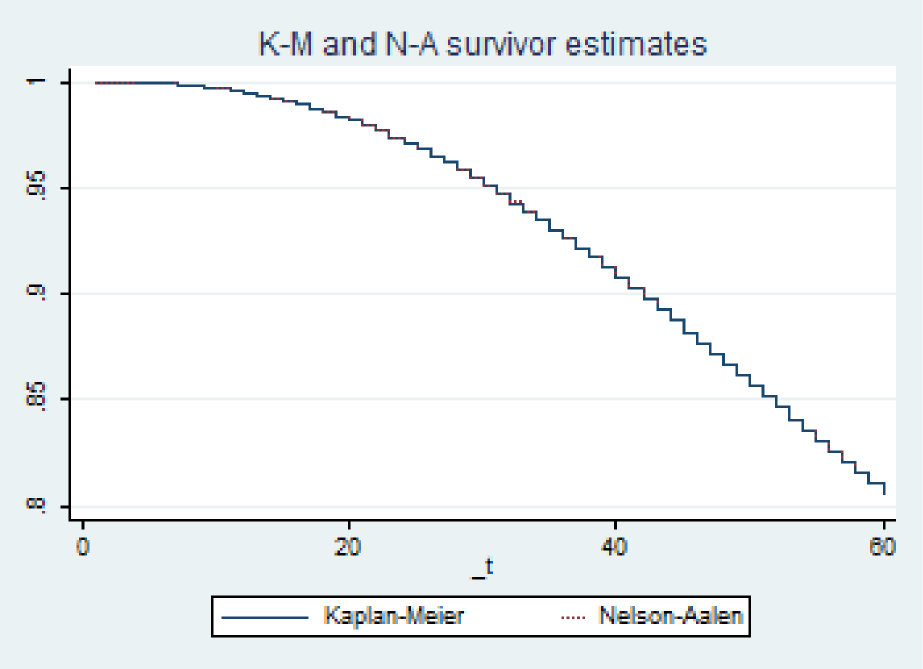
$$\hat{H}(t) = \sum_{j|t_j \leq t} \frac{d_j}{n_j}$$

By the relationship between the survivor and cumulative hazard function, $H(t) = -lnS(t)$, we can get the cumulative hazard function from the Kaplan Meier survivor function, and we also can get the survivor function from the Nelson Aalen cumulative function.

Figure 7.1 shows two nonparametric survivor functions. One comes from the Kaplan Meier survivor function directly, and the other is transformed from the Nelson Aalen cumulative hazard function, according to the relationship between the survivor function and the cumulative hazard function. They are almost the same.

Now we check the survival function by banks. Figure 7.2 shows the survival estimates over time by different bank groups. We focused on the big banks, Chase,

Figure 7.1: Kaplan Meier and Nelson Aalen Survivor Function



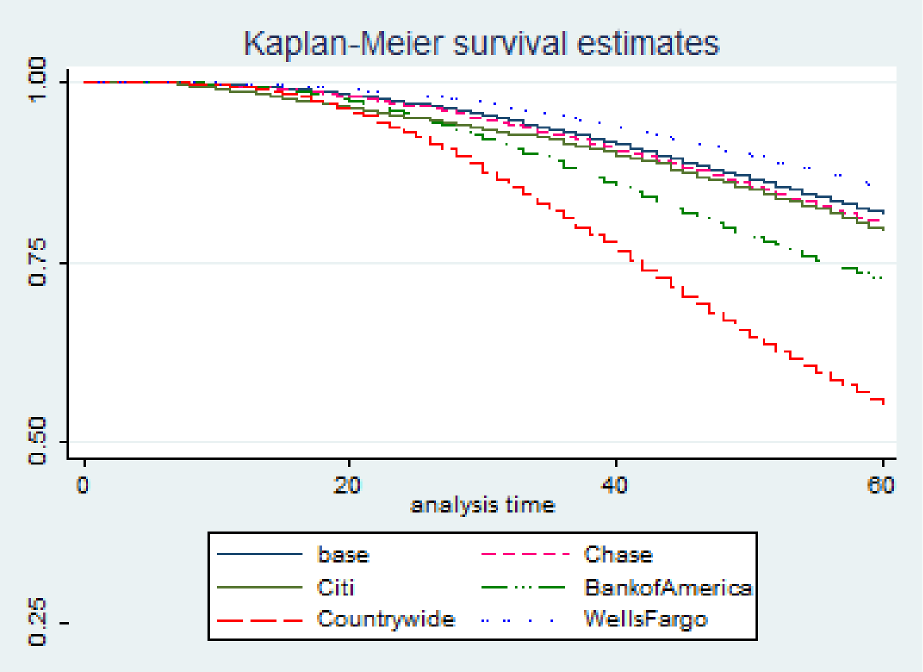
Citi, Bank of America, Countrywide and Wells Fargo, while the remaining loans originated by other smaller banks defined as the base group. Of all the tbanks, Countrywide has the lowest survivor rate and Wells Fargo has the highest survivor rate. Small banks (base group) has a erlatively high survivor rate, greater than all other big banks except Wells Fargo. Chase and Citi are close, with survivor rates higher than Bank of America and Countrywide, but lower than the smaller banks and Wells Fargo. Over time, the survivor rate for all different banks decreases. It is 1 at the beginning, and keeps decreasing over time.

We also estimate the mean of the survival time, defined as

$$\mu_T = \int_0^{t_{max}} \hat{S}(t)dt$$

where t_{max} is the maximum observed failure time. The integral above is restricted to the range $[0, t_{max}]$ because the Kaplan Meier estimator is not defined beyond the largest observed failure time. Therefore, the mean estimated by using the above for-

Figure 7.2: K-M Survivor Function by Bank



mula is often referred to as a restricted mean. A restricted mean μ_T will underestimate the true mean μ_T if the last observed analysis time is censored.

The standard error for the restricted mean is given as follows:

$$\hat{SE}\hat{\mu_T} = \sum_{i=1}^{n} \hat{A}_i \sqrt{\frac{d_i}{R_i(R_i - d_i)}}$$

where the sum is over all distinct failure times, $\hat{A}_i$ is the estimated area under the Kaplan Meier product-limit survivor curve from time t_i to t_{max}, R_i is the number of subjects at risk at a time t_i, and d_i is the number of failures at time t_i.

Table 7.3: Survivor Function by Bank

Bank Group	No. of subjects	Mean	Std. Err.	[95% Conf.	Interval]
Base	274927	56.31(*)	.02	56.26	56.36
Chase	81191	55.99(*)	.04	55.91	56.08
Citi	20099	55.51(*)	.10	55.31	55.71
BankofAmerica	48277	54.13(*)	.07	54.00	54.27
Countrywide	20759	50.48(*)	.11	50.27	50.70
Wells Fargo	173488	57.30(*)	.02	57.25	57.35

(*) largest observed analysis time is censored, mean is underestimated

The estimated mean survival time of different banks are different; the estimated mean of survival time for small banks is 56.31, greater than any big bank except Wells Fargo. The 95% confidence intervals is $(56.26, 56.36)$ for small banks, and does not overlap with the confidence interval of all the other big banks. The estimated mean is similar to the survival graph; overall, the survival function for all groups are different.

To test the equality of survivor function across banks, we apply the log-rank test. The results are shown in Table 7.4.

Table 7.4: Log-rank Equality of Survivor Test

bankgroup	observed	expected
Base	24016	25527.25
Chase	8940	8698.46
Citi	1899	1644.44
BankofAmerica	6280	4139.20
Countrywide	6170	2262.00
Wells Fargo	12074	17107.65
Total	59379	59379.00
chi2(5) = 9520.85		
Pr >chi2 = 0.0000		

We also do several other equality of survivor function tests, and the results from all different equality tests show that the hypothesis of equality across banks is rejected.

Table 7.5: Survivor Function Equality Test

Wilcoxon (Breslow)	Tarone-Ware test	Peto-Peto test
chi2(5) = 8453.05	chi2(5) = 9261.61	chi2(5) = 9531.95
Pr>chi2 = 0.0000	Pr>chi2 = 0.0000	Pr>chi2 = 0.0000

7.2 Applying the Cox Proportion Model

Table 7.7 shows the results from the Cox model when credit score, debt to income ratio, and the loans' originated bank are included. It shows that the hazard rate decreases with credit score, and increases with the debt to income ratio. Chase, Citi,

Table 7.6: Survivor Function by Bank

time	Base	Chase	Citi	BoA	CountryWide	Wells Fargo
1	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
8	0.9983	0.9978	0.9952	0.9986	0.9979	0.9990
15	0.9912	0.9894	0.9777	0.9904	0.9832	0.9951
22	0.9773	0.9742	0.9577	0.9654	0.9456	0.9868
29	0.9567	0.9516	0.9369	0.9268	0.8867	0.9723
36	0.9305	0.9239	0.9159	0.8821	0.8110	0.9518
43	0.8986	0.8919	0.8864	0.8327	0.7287	0.9258
50	0.8649	0.8555	0.8502	0.7865	0.6473	0.8972
57	0.8310	0.8171	0.8121	0.7414	0.5794	0.8672

Bank of America, and Countrywide experience a higher hazard rate than the small banks, while Wells Fargo experiences a lower hazard rate than the small banks.

Table 7.7: Results from Cox Model

Variable	Cox PH model
creditscore	.99***
dti	1.03***
bankgroup	
Chase	1.27***
Citi	1.02***
BankofAmerica	1.96***
Countrywide	2.31***
Wells Fargo	.789***

Table 7.8 shows the Cox proportion assumption test. The global test shows that the null hypothesis is rejected at the 15% level, but not at 5% or 10%, and if we only include the loans in a certain year, the null hypothesis will not be rejected at all.

The survival function and the hazard function from the semiparametric model is similar to the nonparametric analysis, since the nonparametric analysis is completely based on data and lets the data speak for itself. Matches with the nonparametric analysis implies the semi-parametric model fits the data overall. Figure 7.3 shows that the hazard function from the Cox-Proportional model matches the hazard function in Figure 7.4 from the nonparametric model.

The overall model fitness can be evaluated by using the Cox-Snell residuals (see Figure 7.5). If the model fits the data well then the true cumulative hazard function

Table 7.8: Cox Model Test

	rho	chi2	df	Prob>chi2
creditsco	-0.008	0.04	1	0.845
dti	0.037	1.02	1	0.313
base				
Chase	0.090	6.30	1	0.012
Citi	-0.049	1.93	1	0.165
BankofAmerica	0.004	0.01	1	0.911
Countrywide	-0.017	0.25	1	0.621
Wells Fargo	0.048	1.81	1	0.178
hline global test		11.21	7	0.130

Figure 7.3: Hazard Rate Cox Proportional

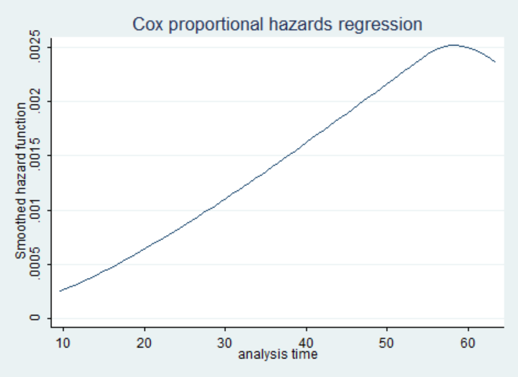
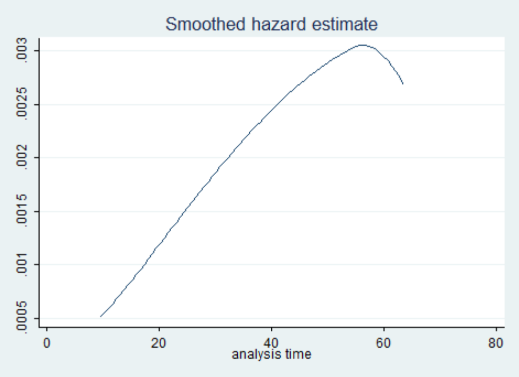
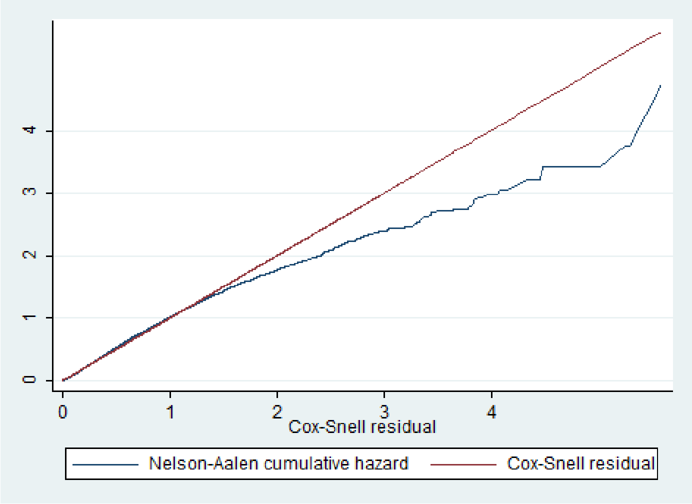


Figure 7.4: Hazard Rate Nonparametric



conditional on the covariate vector has an exponential distribution with a hazard rate of one. We graph the Nelson-Aalen cumulative hazard function and the cs variable so that we can compare the hazard function to the diagonal line. If the hazard function follows the 45 degree line, then we know that it approximately has an exponential distribution with a hazard rate of one and that the model fits the data well. We see that the hazard function follows the 45 degree line very closely, except for very large values of time. It is very common for models with censored data to have some wiggle room at large values of time and it is not something which should cause much concern. It may not be the perfect; overall we would conclude that the final model fits the data well.

Figure 7.5: Cox-Snell Residual vs. NA cumulative Hazard



7.3 Applying a Competing Risk Model

As we have emphasized before, in the situation of competing risk, we should consider the cumulative incidence of a specific event in which we are interested.

$$CIF(ti) = 1 - [1 - CIF_0(ti)]^{exp(x\beta)}$$

where i represent a specific event, (subhazard model). We can estimate the $CIF_0(t)$ and β can be estimated from competing risk models. Then a 3-year cumulative default rate can be written as:

$$CIF(3) = 1 - [1 - CIF_0(3)]^{exp(x\beta)}$$

and similarly for a 5-year cumulative default rate.

$$CIF(5) = 1 - [1 - CIF_0(5)]^{exp(x\beta)}$$

To estimate the base cumulative default rate, we can assume the subhazard rate is the product of the base subhazard rate and the exponential of a linear part of a series of covariates.

$$h(t) = h_0(t)e^{creditscore*\beta_1} \qquad Model\ 1$$

$$h(t) = h_0(t)e^{ccore*\beta_1+DTI*\beta_2} \qquad Model\ 2$$

$$h(t) = h_0(t)e^{cscore*\beta_1+DTI*\beta_2+CLTV*\beta_3} \qquad Model\ 3$$

$$h(t) = h_0(t)e^{cscore*\beta_1+DTI*\beta_2+CLTV*\beta_3+i.oryear*\beta_4} \qquad Model\ 4$$

$$h(t) = h_0(t)e^{cscore*\beta_1+DTI*\beta_2+CLTV*\beta_3+i.oryear*\beta_4+i.bank*\beta_5} \qquad Model\ 5$$

$$h(t) = h_0(t)e^{cscore*\beta_1+DTI*\beta_2+CLTV*\beta_3+i.origyear*\beta_4+i.bank*\beta_5+i.bank*i.oryear*\beta_6} \qquad Model\ 6$$

Model 1 only includes the credit score as the explaining variables; Model 2 includes both credit score and debt to income ratio; Model 3 includes credit score, debt to income ratio, and the cumulative loan to value ratio; Model 4 adds the originated year of loans; Mode 5 adds the dummy of approved bank, Model 6 considers the intersection of loan originated year and approved bank.

Table 7.9 shows how the loan characteristics affect the loan default rate over time for loans originated in Florida by all banks from 2000-2008. In the first model, it only includes the credit score. The second model considers both credit score and the debt to income ratio. The third model considers cumulative loan to value ratio besides all variables in the second model. The fourth models considers the loan originated year besides all variables in the third model. The fifth model consider the loan approved bank besides all the variables in the fourth model, Model 6 included the interaction of the loan originated year and the approved bank.

Table 7.9 shows the results from the competing risk model for loans in Florida originated during 2000-2008. In the first three models, the credit score has the same coefficient of -0.008, which is significant at the 1 % level. DTI has a coefficient of

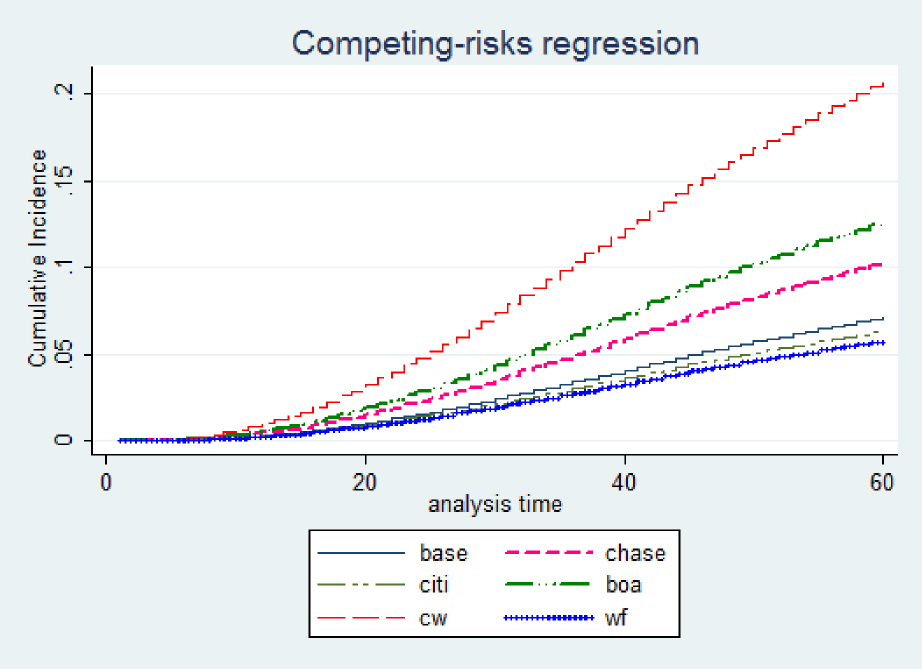
Table 7.9: Estimated Results from Six Models

variable	M1	M2	M3	M4	M5	M6
creditsco	-0.008***	-0.008***	-0.008***	-0.010***	-0.010***	-0.010***
dti		0.026***	0.026***	0.009***	0.010***	0.012***
cltv			0.014***	0.031***	0.031***	0.030***
$oryear$0102				0.000	0.000	
$oryear$0304				1.249***	1.286***	1.160***
$oryear$0506				3.150***	3.160***	3.073***
$oryear$0708				3.364***	3.330***	3.181***
$bankbase$					0.000	
$bankchase$					0.154***	0.013
$bankciti$					0.135	0.262
$bankboa$					0.353***	0.029
$bankcw$					0.134**	1.028*
$bankwf$					-0.163***	-0.454**
$IbanXorchase$0304						0.266
$IbanXorchase$0506						0.023
$IbanXochase$0708						0.181
$IbanXorciti$0304						-0.202
$IbanXorciti$0506						-0.140
$IbanXociti$0708						-0.014
$IbanXorboa$0304						0.461
$IbanXorboa$0506						0.474*
$IbanXoboa$0708						0.476*
$IbanXorcw$0304						-0.787
$IbanXorcw$0506						-1.014*
$IbanXorcw$0708						-0.788
$IbanXorwf$0304						0.277
$IbanXorwf$0506						0.384*
$IbanXowf$0708						0.276

0.026, which implies that when DTI ratio increase, the hazard increases. When the originated year and banks variables are considered in the fifth model, almost all the coefficients are significant. The result in the fourth model shows that loans originated during the 2007-2008 period have the highest hazard rate. The hazard ratio relative with loans originated in 2000 to 2001 is $exp(3.364)$. When the bank variable is considered, we notice that Chase has a hazard ratio of exp(0.154), significant at the 1% level; Citi has a coefficient of 0.135, but not significant; Bank of America has a coefficient of 0.353, significant at 1% level, implying that Bank of America has a hazard probability $exp(0.353)$ higher than small banks. Wells Fargo has a coefficient of -0.163, which is significant and indicates that Wells Fargo has a hazard probability $1 - exp(-0.163)$ lower than the small banks.

Figure 7.6 shows that the predicted cumulative default rate by bank with competing risk model for loans originated in Florida state during 2000-2008. It is shown that countrywide has the highest cumulative default rate, then Bank of America, Chase,

Figure 7.6: Cumulative Incidence by Bank-FG Method



all small banks, then Citi and Wells Fargo.

Table 7.10: Estimated X-year Cumulative Default Rate for Loans in Florida Originated During 2007-2008

bank	CIF(t)	actual	M1	M2	M3	M4	M5	M6
base	cif(3)	17.31	100.00	5.21	4.96	15.64	14.86	14.50
chase	cif(3)	21.03	100.00	4.34	4.07	13.30	14.39	13.83
citi	cif(3)	24.17	100.00	5.28	4.97	16.46	17.56	17.50
boa	cif(3)	18.49	100.00	4.00	3.92	13.10	16.81	18.41
ctryw	cif(3)	20.50	100.00	5.61	5.50	18.31	19.54	20.71
wellsf	cif(3)	15.41	100.00	5.29	5.03	16.18	13.34	12.54
base	cif(5)	26.86	100.00	10.65	10.16	28.53	27.31	26.64
cahse	cif(5)	29.40	100.00	8.93	8.41	24.81	26.64	25.58
citi	cif(5)	32.40	100.00	10.80	10.21	29.94	31.69	31.15
boa	cif(5)	29.93	100.00	8.26	8.08	24.23	30.11	32.38
ctryw	cif(5)	34.83	100.00	11.47	11.26	33.07	34.99	36.62
wellsf	cif(5)	24.35	100.00	10.83	10.32	29.37	24.82	23.39

Source: Freddie Mac Single Family Loan-Level Dataset

Table 7.10 shows the predicted cumulative default rate for loans originated in Florida during the financial crisis. As we can see, when we only include credit score, the predicted 3-year default rate is 1, the model cannot estimate correctly. When we add the debt to income ratio to the model, the 3-year and 5-year default rates can be estimated, but it is far different from the actual default rate for each bank. In the third model, the loan to value ratio is considered and the predicted result still deviates from the actual default rate. However, once the loan originated year is considered, the predicted result improve significantly and matches the actual results better. When banks are considered in the fifth model, the estimated default rate matches the actual default rate very well and the interaction between the bank and time does not improve the estimated result significantly, perhaps it is because when the interaction is considered, many coefficients become insignificant. The fifth model fits the data best, and is consistent with the analysis mentioned right after the model estimated coefficient table.

Table 7.11: Loan Overall Situation vs SettlementS

	Pred	Actu	Diff	No.Loans(99-08)	No.Loans(07-08)	Settlement
Base Group	24.45%	31.37%	6.92%	382,724(47.37%)	53,814(43.29%)	-
Chase	22.28%	32.14%	9.86%	99,120(12.27%)	14,346(11.54%)	29
Citi	23.48%	36.04%	12.57%	24,913(3.08%)	3,248(2.61%)	12
Bankofamerica	21.74%	35.96%	14.22%	62,439(7.73%)	20,647(16.61%)	74
Countrywide	29.12%	42.24%	13.12%	33,430(4.14%)	12,307(9.90%)	-
Wells Fargo	24.47%	27.75%	3.28%	205,368(25.42%)	19,955(16.05%)	10

Table 7.11 shows the overall loan situation of loans in Florida. The first two columns show the predicted and actual 5-year default rates, and the third column shows the difference between the predicted and actual default rate. The greater the number, the worse the mortgage loans management efficiency level. The fourth column shows the number and proportion of loans originated by banks in Florida from 1999-2008. The fifth column is the number and proportion of loans approved by each bank from 2007 to 2008. The last column shows the total settlement amount for each bank.

From the performance index in the third column, we can see that Bank of America and Countrywide have a high value in the difference between the actual and predicted default rate, which reflects a low efficiency level of mortgage loans management. Citi and Chase follow Bank of America and Countrywide. Wells Fargo ranks at the bottom, indicating it has the highest mortgage loan management efficiency.

The fourth column shows the number and proportion of loans by each bank approved from 1999 to 2008. The fifth column shows the number and proportion of loans by each bank approved during the crisis period (2007-2008). Bank of America and Countrywide became more aggressive during crisis period, and took a larger market share during this time. It is not surprising that they have the worst performance index - the difference between actual and predicted default rate. The fact that the banks tightened their mortgage loan policies and approved few loans during the crisis period reflect their conservative attitude.er

Bank of America and Countrywide approved a total of 26% (16.6% and 9.9%) of loans during the crisis period, while Chase approved 11.54% of all loans during this same period, and Citi underwrote only 2.6% of loans for the same period. Bank of America paid a total of $74 billion in settlement, Chase paid a total of $29 billion, Citi and Wells Fargo paid $12 billion and $10 billion. Considering the number of loans and the performance, the settlement amounts are reasonable for most of the banks, except Wells Fargo. Wells Fargo tightened its mortgage loan policy, decreasing its market share during the crisis, and most importantly, has the best performance index, indicating the highest management efficiency level of all banks.

As explained before, during the same period, different banks had different strategies for approving mortgage loans. Some raised the requirement for the applicant's credit score, while other focused on LTV or DTI ratios. The loans approved by different banks have different characteristics in terms of the distribution of credit score, DTI, and LTV ratios. In some cases the same bank changed its policy over years to adapt to the changing environment. It is not surprising then to see that different banks have different loan default rates, and that default rates for the same bank changed over the years, considering the fact that the loan characteristics are different across banks and over time and that the real estate market changed over time.

We also know that different banks have a different market share in each state in the U.S, and different states suffered from the financial crisis differently. To make the default rate across banks comparable, we focused on the mortgage loans approved in the same state, and compared the default rates across banks for loans that were originated during the same time period.

First, from the model results, we see that when loans characteristics are controlled for loans originated in the financial crisis period, all big banks, except Wells Fargo, have positive and significant coefficients, which indicates that they have a higher probability of hazard than loans by small banks, when loans characteristics are controlled. Wells Fargo had a negative and significant coefficient, which indicates that Wells Fargo did a better job than small banks in terms of loan management efficiency.

Second, from the difference between actual and predicted loan default rates, we can see how banks are different in the efficiency level of mortgage loans management. Bank of America and Countrywide performed the worst; Citi and Chase slightly better than small banks. Wells Fargo did better than the small banks.

In addition, we have mentioned how missing data in the data section affected the results. Although only 2.5% of the data is missing, it provides a good perspective as to whether there was problem. In general, the loans with missing data should have been considered low credit loans, and expected to default more frequently than loans with complete data. As was pointed out earlier, Bank of America and Countrywide have a lower default probability for loans with missing data than loans with complete data. This is inconsistent with other banks and with our intuition.

Provided with the different number/proportion of loans approved by different banks during the financial crisis, we can see that there is a solid ground for a different level of settlement with each bank.

Overall, there are significant reasons for big banks to agree to the huge settlements. Although the information is not publicly available, we find evidence from the public data that Bank of America and Countrywide did misrepresent the quality of their loans, from both the missing data and non-missing data loan performance. The loans by Bank of America and Countrywide have the highest hazard rate over time and the lowest survival rate over time when all other factors are included. Bank of America and Countrywide also performed much worse than their predicted default rate during the financial crisis. This explains the $74 billion settlement, the largest penalty levied against big banks by the federal government. Part of the reason Bank of America and Countrywide did poorly is the good work they did in earlier years, which led to their over-confidence later in underestimating loan risk, especially when the market downturn. This does not justify their wrongdoing, since the loans they originated performed much worse than predicted.

Chase and Citi performed worse than expected, but not worse than other banks, which explains why their settlements are less than Bank of America. Citi and Chase did poorly in the earlier years, and were more cautious in making later loans, and so did not fare as badly. Wells Fargo did the best out of all big banks, their coefficients in all models are significantly better, and this is not coincidental.

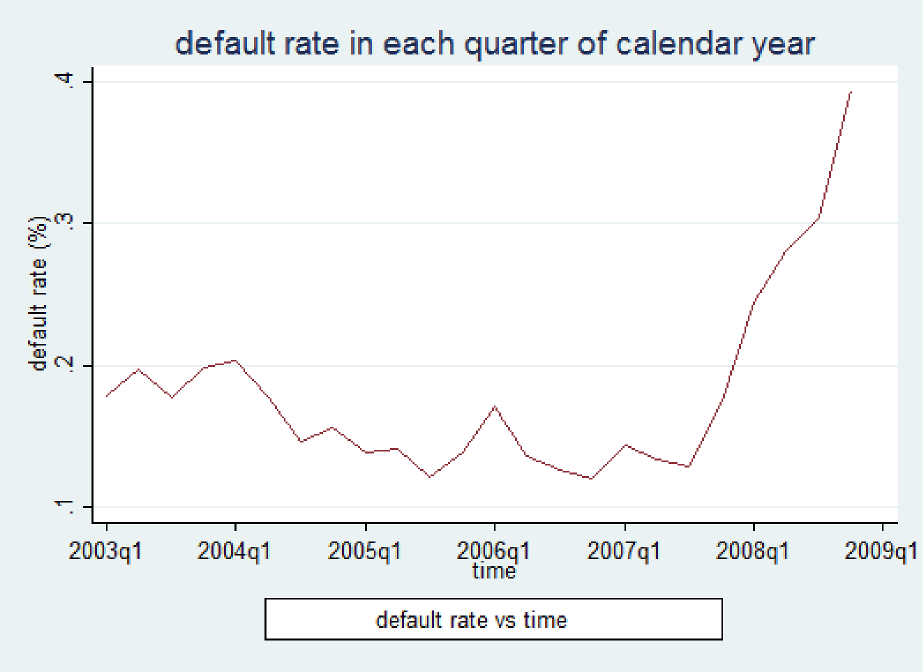
Above all, regulators made reasonable settlement deals with most banks; although from this perspective, justice seems based on the amount of settlement. While Wells Fargo did the best of all the big banks, it was not accidental, since during 2001-2005, Wells Fargo was continuously accused (and punished) for a number of instances of alleged wrongdoing by the SEC and other regulators. Had they been found guilty of illegal activities, they would have had to pay fines, which is perhaps why they kept a cautious eye on all their later loans, which performed much better than the industry average. If a bank like Wells Fargo could perform well during the financial crisis, and still get punished, then we might well need to rethink what justice means in a free society. Wells Fargo continued to battle FHA to avoid further punishment as late as May 2014, and failed. Although it failed in June, this sends a message that if the bank believes in its innocence, and does not deserve such severe punitive measures, it will not give up the legal battle or easily settle for such an unfair agreement. For now we can only wait and see what the final outcome will be.

CHAPTER VIII

Policy Implications

8.1 Default Rate over Time

Figure 8.1: Default Rate for Loans in Each Quarter of Calendar Years



The default rate affects policy, and policy further affects the default rate. If we follow the default rate in each month in each calendar year, then we can rely on the trend of the default rate and make better policy decisions. Table 8.1 shows how loan default rates varies over time. The default rate in each quarter of the calendar year as the actual default rate during that quarter over the cumulative active loans (excluding loans already defaulted and prepaid) originated before that quarter.

From Figure 8.1, we can see that the default rate varied a little from 2003 to 2004. During 2004 to the second quarter of 2007, the default rate overall followed a decreasing trend, except for a small peak in the first quarter of 2006. Beginning with the second quarter of 2007, the default rate increases, and by the second quarter of 2008 it has almost doubled from the second quarter of 2007. it continued to increase after that.

8.2 The Number of Loans Originated over Time

Table 8.1: The Number of Loans Originated over Time

	Chase	Citi	BOA	CW	WF	Others	All
Jan-07	804	177	889	1,950	1,398	4,944	10,162
Feb-07	690	168	1,034	1,631	1,347	4,347	9,217
Mar-07	686	201	1,050	1,096	1,257	4,191	8,481
Apr-07	772	300	1,060	695	1,335	4,305	8,467
May-07	426	428	683	851	1,234	4,001	7,623
Jun-07	255	484	701	763	1,292	4,372	7,867
Jul-07	568	350	851	351	1,112	3,908	7,140
Aug-07	910	532	797	529	1,389	4,610	8,767
Sep-07	863	464	587	393	1,273	4,629	8,209
Oct-07	1,267	439	870	1,398	1,884	5,770	11,628
Nov-07	1,265	794	858	1,106	1,992	5,108	11,123
Dec-07	1,686	276	559	1,158	1,775	5,657	11,111
Jan-08	1,483	139	554	718	1,487	4,854	9,235
Feb-08	1,110	457	753	731	1,514	4,209	8,774
Mar-08	746	668	520	546	1,184	3,678	7,342
Apr-08	333	280	315	577	707	2,706	4,918
May-08	530	412	355	437	1,250	2,753	5,737
Jun-08	423	414	405	481	1,344	3,139	6,206

Source: Freddie Mac Single Family Loan-Level Dataset

The relative stable and low default rate in the fist half year of 2007 led banks to approve a relatively large number of loans during the second half of 2007. As we noted above, the default rate increased dramatically after the third quarter of

2007. This trend led to banks to tighten their mortgage loan policies again. Chase started to tighten its policy beginning in January 2008; Citi tightened its policy in December 2007, Bank of America and Countrywide tightened their policies starting in November 2007; Wells Fargo tightened its polcy from December 2007.

8.3 Final Thoughts

Many factors can affect the loan default rate. From the discussion above, we already see there is a significant relationship between loan characteristics and the loan default rate: loan default rates increase as DTI and LTV ratio increases, and good credit score can significantly decrease the default risk of loans. The loan default rate also changes with the originated year because the policy of underwriting loans changes over time. If the overall mortgage policy tightens, with all other variables the same except for stricter loan processing policy, we can expect the default rate to decrease. On the other hand, if all other variables remain the same, but the regulators loosen the policy, with all other conditions the same, we can expect the default rate to increase.

However, we should also keep in mind that loan characteristics themselves cannot explain the whole story of the default rate. The default rate is not only affected by loan characteristics, but by the overall macroeconomic situation, from the GDP, consumption index, to the unemployment rate. To make the right policy decisions, we should consider all kinds of factors, from macro perspective to micro perspective.

Considering the default rate and the change in the number of loans originated by banks over time, we think that the banks also kept following the default rate and adjusted their loan policies accordingly. In other words, the banks' loan policies already reflected the risk change over time. However, under the current mechanism, some systematic risk just could not be avoided easily.

To avoid the systematic risk, we probably need to set a completely different market mechanism. For example, with the current mechanism, the interest rate is determined by the supply and demand of mortgage backed security in the market. However, as the price of loans, the interest rate should reflect the loans' risk that is involved. Another factor we need to consider is the separation between loan originator (bank)and loan owner (some banks are owners, but the majority are owned by Fannie and Freddie.)

If we can overcome these problems, then the risk could be significantly controlled. Otherwise, we can only avoid unsystematic risk and trying to make the right policy decisions based with limited choices.

BIBLIOGRAPHY

BIBLIOGRAPHY

(2014), The criminalisation of American business; Corporate settlements in the United States., *The Economist*, (8902), 9.

Aalen, O. (1978), Nonparametric inference for a family of counting processes, *The Annals of Statistics*, pp. 701–726.

Agarwal, S., B. W. Ambrose, S. Chomsisengphet, and A. B. Sanders (2012), Thy neighbors mortgage: Does living in a subprime neighborhood affect ones probability of default?, *Real Estate Economics*, *40*(1), 1–22.

Altman, E. I. (1968), Financial ratios, discriminant analysis and the prediction of corporate bankruptcy, *The Journal of Finance*, *23*(4), 589–609.

Backman, M. (2014), Bank of America to pay $16.65 billion over mortgages, [Online:Cnn.com, posted August 21, 2014].

Banasik, J., J. N. Crook, and L. C. Thomas (1999), Not if but when will borrowers default, *Journal of the Operational Research Society*, pp. 1185–1190.

Been, V., M. Weselcouch, I. Voicu, and S. Murff (2013), Determinants of the incidence of us mortgage loan modifications, *Journal of Banking & Finance*, *37*(10), 3951–3973.

Bonilla, D. (2015), Zombie houses on Long Island, [Online:newsday.com, posted 2015].

Breslow, N. (1970), A generalized kruskal-wallis test for comparing k samples subject to unequal patterns of censorship, *Biometrika*, *57*(3), 579–594.

Breslow, N. (1974), Covariance analysis of censored survival data, *Biometrics*, pp. 89–99.

Chan, S., M. Gedal, V. Been, and A. Haughwout (2013), The role of neighborhood characteristics in mortgage default risk: Evidence from new york city, *Journal of Housing Economics*, *22*(2), 100–118.

Chiang, R. C., Y.-F. Chow, and M. Liu (2002), Residential mortgage lending and borrower risk: the relationship between mortgage spreads and individual characteristics, *The Journal of Real Estate Finance and Economics*, *25*(1), 5–32.

Cleves, M. (2000), Analysis of multiple failure-time data with stata, *Stata Technical Bulletin*, *9*(49).

Cleves, M. (2008), *An introduction to survival analysis using Stata*, Stata Press.

Coviello, V., and M. Boggess (2004), Cumulative incidence estimation in the presence of competing risks, *Stata Journal*, *4*, 103–112.

Cox, D. R. (1972), Regression models and life-tables, *Journal of the Royal Statistical Society. Series B (Methodological)*, pp. 187–220.

Deng, Y., A. D. Pavlov, and L. Yang (2005), Spatial heterogeneity in mortgage terminations by refinance, sale and default, *Real Estate Economics*, *33*(4), 739–764.

Desai, V. S., D. G. Conway, J. N. Crook, and G. A. Overstreet (1997), Credit-scoring models in the credit-union environment using neural networks and genetic algorithms, *IMA Journal of Management Mathematics*, *8*(4), 323–346.

Donald, E. R., L. Kartono, and H. Richard (1996), Borrower risk signaling using loan-to-value ratios, *Journal of Real Estate Research*, *11*(1), 71–86.

Dungey, D. (2007), Ficos and aus: We will add your distinctiveness to our collective, *Calculated Risk blog: The Compleat UberNerd*.

Efron, B. (1977), The efficiency of cox's likelihood function for censored data, *Journal of the American statistical Association*, *72*(359), 557–565.

Fine, J. P., and R. J. Gray (1999), A proportional hazards model for the subdistribution of a competing risk, *Journal of the American statistical association*, *94*(446), 496–509.

Fitch, S. (2010), No-doc mortgages are back, [Online:Forbes.com, posted July 02, 2010].

Gehan, E. A. (1965), A generalized wilcoxon test for comparing arbitrarily singly-censored samples, *Biometrika*, *52*(1-2), 203–223.

Gerardi, K., L. Goette, and S. Meier (2013), Numerical ability predicts mortgage default, *Proceedings of the National Academy of Sciences*, *110*(28), 11,267–11,271.

Grambsch, P. M., and T. M. Therneau (1994), Proportional hazards tests and diagnostics based on weighted residuals, *Biometrika*, *81*(3), 515–526.

Greene, W. H. (1996), Marginal effects in the bivariate probit model.

Hall, W. J., and J. A. Wellner (1980), Confidence bands for a survival curve from censored data, *Biometrika*, *67*(1), 133–143.

Hess, K. R. (1995), Graphical methods for assessing violations of the proportional hazards assumption in cox regression, *Statistics in medicine*, *14*(15), 1707–1723.

Holmes, S. (1999), Fannie mae eases credit to aid mortgage lending, *New York Times*.

Hua, Z., Y. Wang, X. Xu, B. Zhang, and L. Liang (2007), Predicting corporate financial distress based on integration of support vector machine and logistic regression, *Expert Systems with Applications*, *33*(2), 434–440.

Irvine, C. (2015), 1.1 million u.s. properties with foreclosure filings in 2014, down 18 percent from 2013 to lowest level since 2006, [Online:Realtytrac.com; posted January 15, 2015].

Isidore, C. (2013), Bank of America in $10 billion settlement with Fannie Mae, [Online:Cnn.com, posted January 07, 2013].

Jacobson, T., and K. Roszbach (2003), Bank lending policy, credit scoring and value-at-risk, *Journal of banking & finance*, *27*(4), 615–633.

Kau, J. B., and D. C. Keenan (1999), Patterns of rational default, *Regional Science and Urban Economics*, *29*(6), 765–785.

Kau, J. B., D. C. Keenan, and H. J. Munneke (2012), Racial discrimination and mortgage lending, *The Journal of Real Estate Finance and Economics*, *45*(2), 289–304.

Laitinen, E. K., and T. Laitinen (2001), Bankruptcy prediction: application of the taylor's expansion in logistic regression, *International Review of Financial Analysis*, *9*(4), 327–349.

Lambrecht, B., W. R. Perraudin, and S. Satchell (2003), Mortgage default and possession under recourse: A competing hazards approach, *Journal of Money, Credit, and Banking*, *35*(3), 425–442.

Lawrence, E. C., L. D. Smith, and M. Rhoades (1992), An analysis of default risk in mobile home credit, *Journal of Banking & Finance*, *16*(2), 299–312.

Lin, T. T., C.-C. Lee, and C.-H. Chen (2011), Impacts of the borrower's attributes, loan contract contents, and collateral characteristics on mortgage loan default, *The Service Industries Journal*, *31*(9), 1385–1404.

Lunn, M., and D. McNeil (1995), Applying cox regression to competing risks, *Biometrics*, pp. 524–532.

Mantel, N., and W. Haenszel (1959), Statistical aspects of the analysis of data from retrospective studies, *J natl cancer inst*, *22*(4), 719–748.

Marubini, E., and M. G. Valsecchi (2004), *Analysing survival data from clinical trials and observational studies*, vol. 15, John Wiley & Sons.

Narain, B. (1992), Survival analysis and the credit granting decision, *Credit scoring and credit control*, pp. 109–122.

Nelson, W. (1972), Theory and applications of hazard plotting for censored failure data, *Technometrics*, *14*(4), 945–966.

Peto, R., and J. Peto (1972), Asymptotically efficient rank invariant test procedures, *Journal of the Royal Statistical Society. Series A (General)*, pp. 185–207.

Premachandra, I., G. S. Bhabra, and T. Sueyoshi (2009), Dea as a tool for bankruptcy assessment: A comparative study with logistic regression technique, *European Journal of Operational Research*, *193*(2), 412–424.

Prentice, R. L. (1978), Linear rank tests with right censored data, *Biometrika*, *65*(1), 167–179.

Rexrode, C., and A. Grossman (2014), Record bank of America settlement latest in government crusade., *Wall Street Journal (Online)*, p. 1.

Schoenfeld, D. (1982), Partial residuals for the proportional hazards regression model, *Biometrika*, *69*(1), 239–241.

Setzer, G. (2008), Fannie mae tightens loan criteria for credit scores, *Mortgagenewsdaily.com*.

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

Tarone, R. E., and J. Ware (1977), On distribution-free tests for equality of survival distributions, *Biometrika*, *64*(1), 156–160.

Tsai, C.-F., and J.-W. Wu (2008), Using neural network ensembles for bankruptcy prediction and credit scoring, *Expert Systems with Applications*, *34*(4), 2639–2649.

Watkins, T., S. Valley, and T. Alley (2009), The nature and the origin of the subprime mortgage crisis.

West, D. (2000), Neural network credit scoring models, *Computers & Operations Research*, *27*(11), 1131–1152.