Who defaults on their home mortgage?

Doviak, Eric and MacDonald, Sean

24. September 2011

Who Defaults on their Home Mortgage?

Eric Doviak* Sean MacDonald†

September 24, 2011

Abstract

Since February 2010, detailed information on every home mortgage default and foreclosure in New York State must be filed with the New York State Banking Department (NYSBD). Pairing the NYSBD's data with data on originations from the Home Mortgage Disclosure Act (HMDA) enables us to identify the race and ethnicity of borrowers who defaulted on their home mortgages (in New York State).

Like many previous studies, we find strong racial and ethnic disparities in lending practices, but we do not find conclusive evidence that HMDA-measurable forms of discrimination increased a borrower's probability of default. After controlling for other factors, we find that the interest rates charged to black and Latino borrowers tended to be higher than the ones charged to their white and non-Latino counterparts. This may be one reason why blacks and Latinos tend to default at a higher rate, but other factors, such as the tendency of black and Latino borrowers to take out larger loans than their white and non-Latino counterparts, may also have contributed to the higher default rate among black and Latino borrowers.

*Brooklyn College, City University of New York – eric@doviak.net

†New York City College of Technology, City University of New York – smacdonald@citytech.cuny.edu

1 Introduction

Under a state law that was enacted on December 15, 2009, mortgage servicers must send a "pre-foreclosure filing" (PFF) notice to delinquent borrowers at least 90 days prior to filing for foreclosure on a primary residence in the State of New York. The notice informs homeowners that their loan is in default, lists the amount necessary to cure the default and lists measures that they can take to avoid foreclosure, such as negotiating a loan modification with their lender and consulting with a non-profit housing counselor (New York State Banking Department, 2009).

Since February 13, 2010, mortgage servicers are also required to file the notices with the New York State Banking Department (NYSBD), which collected an extraordinary level of detail on the loans. Among the many data fields collected are: the property address, the names of the borrowers, the current monthly payment, the delinquent contractual payments, the interest rate, whether the loan is a fixed-rate or adjustable-rate mortgage, the date and the amount of the original loan, the lien type, the loan term, whether the loan has been modified or not and whether an investor's approval is necessary to modify the loan. If the loan progresses to a *lis pendens* filing (i.e. the first step in the foreclosure process – the filing of the complaint), then servicers are also required to follow up on their initial filing with information on the entity filing for foreclosure.

The detail captured in the PFF data makes three forms of analysis possible. First, we can match the defaulted loans to publicly available data on originations from the Home Mortgage Disclosure Act (HMDA). By combining the HMDA and PFF data, we can see which borrowers were more likely to default. Second, we can compare the loans that entered the foreclosure process to those that did not. Third, we can use the "Full PFF" dataset to compare defaulted loans across the years in which they were originated.

The first two datasets – which compare originations to defaults and compare defaults to lis pendens filings – effectively generate a quasi-longitudinal analysis, in the sense that we track the universe of New York State home mortgages from origination to default to foreclosure. We use the term "quasi-longitudinal" however, because the PFF data only provides information on borrowers who defaulted in 2010. We do not have data on the borrowers who did not default or defaulted in other years. Moreover, there is no way to perfectly match the PFF data to the HMDA data. We believe that our matching strategy generates a reasonably accurate result, but there is no way to verify the match.

This paper discusses the findings of our analysis of the combined HMDA and PFF datasets. In the interest of brevity and to avoid confusion between the two datasets, we discuss our analysis of the PFF dataset (which tracks loans from default to foreclosure) in a separate paper (Doviak and MacDonald, 2011).

After matching the PFF data to the 2004-2008 HMDA data, we found that the loan amount was the best predictor of default. This finding is not surprising. All else equal, a borrower who took out a larger loan was left with less equity (or even negative equity) in his/her home after home prices tumbled during the recent economic and financial crisis. Consequently, the borrower who took out a larger loan may have had greater incentive to walk away from the loan and shift the loss onto the lender or may have faced greater obtacles in obtaining a loan modification.

What may be surprising to some readers is the extent to which blacks and Latinos borrowed more than their white and non-Latino counterparts. As discussed in section 4, about half of white borrowers and half of non-Latino borrowers borrowed less than $200,000, whereas about a quarter of blacks and a quarter of Latinos borrowed less than $200,000. This is one important reason why the foreclosure crisis has had a disproportionately larger impact on black and Latino communities.

Another interesting finding from the combined HMDA-PFF data is that "middle-income" borrowers (those with incomes between $80,000 and $200,000[1]) seem to have received disproportionately more pre-foreclosure filing notices than low-income and high-income borrowers. Section 4 will also show that black and Latino borrowers were disproportionately middle-income, yielding another reason why the foreclosure crisis has had a disproportionately larger effect on their communities.

The combined HMDA-PFF data also shows that borrowers who took out "high-cost" loans (i.e. loans in which the interest rate at origination is three or more percentage points above the yield on the comparable U.S. Treasury bond) were more likely to default. This is not surprising. One would expect lenders to charge a higher rate to borrowers who are more likely to default as compensation for the additional risk. Moreover, a higher interest rate increases the borrower's monthly payments, thus making the loan more difficult to repay.

What may be surprising to the reader is the extent to which blacks and Latinos were more likely to take out high-cost loans. This is a third reason why the foreclosure crisis has had a disproportionately larger effect on black and Latino communities.

Previous studies have argued that blacks and Latinos faced discriminatory lending practices. However, we must first check to see if there some were fundamental factors which help explain why they were more likely to take out high-cost loans.

To perform such a check, we controlled for other factors by regressing their loan's rate spread (i.e. the difference between the interest rate and the yield on the comparable U.S. Treasury bond) on a large number of variables, such as income, loan amount, whether there was a co-borrower on the loan, the purpose of the loan, etc. and dummy variables for race and ethnicity. As discussed in section 5, the regressions suggest that the interest rates on loans originated to blacks and Latinos were higher than those originated to their white and non-Latino counterparts and that the differences were statistically significant. The HMDA data does not contain important variables (such as the borrower's credit score and the loan-to-value) however, so the evidence of discrimination is not conclusive.

Assuming that blacks and Latinos did face discrimination however, the question of whether that discrimination increased their probability of default remains open. So in a second step, we regressed the probability of default on the predicted rate spread, a large number of other variables and dummy variables for race and ethnicity. In those regressions, the coefficient on the predicted rate spread was not statistically significant from zero in one of the regressions, but the coefficients on the dummy variables for black race and Latino ethnicity were positive and statistically significant.

Consequently, we cannot conclude that the HMDA-measurable form of discrimination (i.e. the rate

[1]This odd definition of "middle-income" corresponds to the income brackets that had a greater than average rate of default.

spread) increased their probability of default. Instead, we have to accept at face value the finding of the regression analysis – that blacks and Latinos default at a significantly higher rate after controlling for other factors.

We do not believe however that the melanin level in a person's skin makes him/her more or less likely to default on his/her home mortgage. Instead we believe that black race and Latino ethnicity must be acting as a proxy for some missing variable that does increase their probability of default, such as differences in socio-economic characteristics, racial and ethnic disparities in the effect of the recent economic recession and/or forms of discrimination that we cannot measure with the HMDA data.

Because the foreclosure crisis has had a disproportionately large effect on blacks and Latinos, one cannot understand the mortgage foreclosure crisis without understanding its racial and ethnic dimensions. For this reason, we discuss the literature on discrimination in mortgage lending in section 2 and we will discuss the racial and ethnic disparities that we observe in the data in section 4.

In section 3, we describe the PFF data in more detail and explain how we paired it with the HMDA data. Section 4 compares the loan origination data to the pre-foreclosure filing data and attempts to explain why blacks and Latinos tend to default at a higher rate than their white and non-Latino counterparts. Section 5 provides a very basic regression analysis that attempts to resolve some of the puzzles that we find in the comparisons and continues to explore the racial and ethnic dimensions of the foreclosure crisis. Section 6 concludes.

2 Review of the Discrimination Literature

The subprime mortgage foreclosure crisis, which began to unfold in 2006, was one of the first indicators of the forthcoming bursting of the nation's housing bubble. It was also perhaps the first sign that the risky home mortgages that formed the basis of subprime lending were the major culprit in the rapidly escalating delinquency and foreclosure rates leading up to the banking and financial crisis of 2008.

To provide some perspective, subprime lending was virtually non-existent at the peak of the previous real estate boom in 1989-90. Subprime loans then increased from 5 percent of total mortgage originations in the U.S. in 1994 to almost 20 percent in 2005 (Doms et al., 2007).

A number of recent studies have examined the relationship between the subprime foreclosure crisis, predatory lending and the concentration of subprime loans within predominantly black and Latino and other minority communities. These studies have sought explore the underlying causes of the subprime foreclosure crisis, from residential segregation within many of the nation's metropolitan areas (Rugh and Massey, 2010), the lack of alternatives to subprime lenders in predominantly minority communities (U.S. Dept. of Housing and Urban Development, 2000b), changes in home prices and home price volatility (Doms et al., 2007) and the inability to exclude the mortgage on a primary residence from protection following the passage of the Bankruptcy Reform Act of 2005 (Morgan et al., 2011).

As mentioned previously, this review focuses on the analyses that explore the racial and ethnic dimensions of high-cost lending and whether discriminatory lending practices contributed to the subprime mortgage foreclosure crisis. We acknowledge that discrimination may have occurred, but the empirical evidence of discrimination suffers from the limitations of the HMDA data, which omits important variables, such as the borrower's credit score, the borrower's other assets and the loan-to-value ratio.

Consequently, it is easy to show that high-cost lending was most prevalent in predominantly minority communities, but it is difficult to take the next step and use the HMDA data to show that such lending is evidence of discrimination.

Bocian et al. (2006) does overcome the some of the HMDA data's limitations however. After pairing the 2004 HMDA data with a proprietary dataset of 177,000 subprime loans, they found that blacks and Latinos received a disproportionate share of high-cost loans after controlling for other factors, including the borrower's FICO score and the loan-to-value ratio. The major limitation of their study however is the fact that it does not contain the universe of originations. Instead the study focuses on the dataset of subprime loans that was available to them, so it is not generalizable to the broader market and it does not explain why borrowers took a subprime loan as opposed to a prime loan.

When employing the HMDA data to study the broader market researchers are generally confined to finding a correlation between racial segregation and the probability of receiving a high-cost loan. For example, Squires et al. (2009) use the 2000 Census data to construct a dissimilarity index to obtain a measure of the ten most segregated and the ten least segregated metropolitan areas in the U.S. They then compare the indices derived to the percentage of high-cost loans originated. Using 2006 HMDA data and the 2006 American Community Survey, they employ a multivariate OLS model (to control for several MSA-level variables) and find that racial segregation is a significant predictor of the percentage of high-cost loan originations in an MSA. Their results suggest that a 10 percent increase in black segregation was associated with a 1.4 percent increase in high-cost loans.

Other studies have also found a link between the racial composition of a neighborhood and its share of subprime lending in that neighborhood. For example, in a joint study conducted by several community organizations, Bromley et al. (2008) focused on subprime lending activity across seven large metropolitan areas in the U.S. in 2006. Data on the number of high-risk loan originations conducted by a sample of 35 subprime lenders during that year indicated that these lenders accounted for an estimated 20 percent of the market share of subprime loans in predominantly minority neighborhoods within these metropolitan areas. Further, more than 40 percent of the loans made by high-risk lenders in these metropolitan areas were in neighborhoods where the share of minority residents was 80 percent or more. Subprime lenders' market share was also positively correlated with the percentage of minority residents within a given census tract.

The U.S. Dept. of Housing and Urban Development (2000b) also finds a disproportionate concentration of subprime lending in predominantly minority – and particularly – African-American communities. In their study, which focuses primarily on subprime refinance lending, the number of subprime refinance loans originated in the New York metropolitan area between 1993 and 1998 increased by an estimated 350 percent. The study also found that subprime loans were three times more likely to be originated in lower-income neighborhoods in the New York metropolitan area than in higher-income neighborhoods, and more than four times more likely in predominantly black than in predominantly white neighborhoods.

It's particularly interesting to note that their study was published in 2000, which indicates that subprime lending expanded rapidly into minority communities long before the subprime mortgage meltdown began in 2006. According to Laderman (2001), one factor which contributed to the expansion of subprime mortgage lending in the early 1990s was the increasing frequency with which mortgages were securitized. Securitization reduced the risk associated with lending to subprime borrowers and it enabled large sums to be assembled for the purpose of subprime lending. Another factor that Laderman cites was deregulation. Prior to passage of the Depository Institutions Deregulation and Monetary Control Act in 1980, limits were imposed on the interest rates that lenders could charge. Once the caps were lifted, lenders could raise interest rates high enough to absorb the risk associated with lending to subprime borrowers.

In a separate but related report, the U.S. Dept. of Housing and Urban Development (2000a) found that the pattern of originating subprime loans to minorities transcended income level and that this pattern established itself long before the subprime loan market reached its peak during the early 2000s. Instead, borrowers in high-income black neighborhoods were twice as likely as those in low-income white neighborhoods to take out a subprime loan. Specifically, the study found that just six percent of borrowers in high-income white neighborhoods had subprime loans while 39 percent of borrowers in upper-income black neighborhoods had subprime loans. This figure was more than twice the 18 percent rate for borrowers in low-income white neighborhoods.

Such findings are disturbing. The lack of information on credit scores in the HMDA data may explain some of the disparities in the rate spreads among individual borrowers, but it is hard to see how this could be applicable across neighborhoods. In other words, it is easy to imagine individual cases where a high-income black borrower's credit score is lower than a low-income white borrower's credit score; however it is difficult to see how the average credit score of a high-income black neighborhood could be lower than the average credit score of a low-income white neighborhood.

Given that blacks and Latinos took a disproportionately high share of subprime loans, one would also expect a disproportionately high rate of foreclosure in black and Latino communities. This is precisely what two other studies have found.

Rugh and Massey (2010) attempt to link the correlation between the high-cost lending and the patterns of residential segregation to the subprime foreclosure crisis. To find the link, they obtained the total number of foreclosures between 2006 and 2008 from RealtyTrac's foreclosure database and computed the foreclosure rate as the number of filings per household unit. They then used the 2004-2006 HMDA data to compute the share of high-cost loans[2] in each MSA. To derive a measure of regulatory oversight, they also computed the share of loans within the MSAs that were originated by institutions covered under the Community Reinvestment Act (CRA).

Employing an OLS multiple regression model, Rugh and Massey regress the number and rate of foreclosures in the nation's 100 largest MSAs on two measures of segregation: residential unevenness and spatial isolation. Their regression results suggest that residential segregation and the share of high-cost loans are both positively correlated with the number and rate of foreclosures across U.S. metropolitan areas.

[2]Rugh and Massey use the term "subprime" to describe high-cost loans.

One frustrating omission in their published paper however is the lack of a regression of the high-cost lending share on measures of racial and ethnic segregation. If segregation enabled lenders to target minorities for high-cost loans (as Rugh and Massey claim), then they should have regressed the high-cost lending share on measures of segregation. If the coefficient were positive and statistically significant, then their claims of racial and ethnic targeting would have a firmer foundation.

Gerardi and Willen (2008) also examine the relationship between foreclosures and subprime lending in urban and minority communities. By matching the 1998-2006 HMDA data to deed registry data in the State of Massachusetts, they generate a dataset that contains the universe of mortgages, foreclosures, purchases and sales. In their analysis of the data, they find that a disproportionate share of subprime loans were originated to blacks and Latinos, but these loans proved unsustainable when home prices fell. The records of property sales in their dataset indicate that a "sudden and severe fall in the share of minority home ownership" began in 2005 due to a significant increase in foreclosures among minority homeowners.

The studies reviewed above show that blacks and Latinos took a disproportionately high share of high-cost and subprime loans, but the evidence that this trend reflects discrimination suffers from the limitations of the HMDA data. Nonetheless, the studies do help explain our finding that blacks and Latinos defaulted on their mortgages at a higher rate than their white and non-Latino counterparts.

3 The New York State Pre-Foreclosure Filing Data

As mentioned in the introduction, in February 2010, the New York State Banking Department (NYSBD) began collecting data on home mortgages in default. When the borrower defaults on his/her primary residence, his/her mortgage servicer sends him/her a "pre-foreclosure filing" (PFF) notice and transmits an extraordinary level of detail on the mortgage to the NYSBD. If the borrower does not cure the default within 90 days, the servicer may commence the foreclosure process with a lis pendens filing. If it chooses to do so, it must also inform the NYSBD of the lis pendens filing.

Given our discussion of the racial and ethnic disparities in mortgage lending during the previous decade, one interesting way to analyze the PFF data is to match the loans to the HMDA originations data and compare the borrowers who defaulted in 2010 to those who did not default.

Because the PFF dataset contains information on both defaults and foreclosures, another interesting way to analyze the data is to compare the defaulted loans that did not progress to foreclosure to those that did. As mentioned previously, we discuss these comparisons in a separate paper (Doviak and MacDonald, 2011).

Prior to comparing originations to defaults however, we first explain how we prepared the PFF dataset for statistical analysis in subsection 3.1. Then, in subsection 3.2, we explain how we matched the PFF data to the HMDA data. After providing those explanations, we discuss our comparisons in section 4 and we provide a very basic regression analysis in section 5.

3.1 Preparing the Data for Analysis

Prior to performing an analysis of the PFF data, we had to remove duplicate filings because servicers who missed the three-business day deadline or submitted incorrect information would "re-file" the loan. Some servicers also submitted one filing for each borrower on the loan.

The duplicates were fairly easy to identify however, because servicers almost always included their loan numbers with the filing, so the combination of the servicer's identity and the loan number enabled us to uniquely identify each loan[3]. In cases where a servicer submitted one filing for each borrower, we compared the borrower's first and last name to the names of other borrowers on the loan to see if there was a co-applicant or not.

Because servicers re-filed a loan to correct mistakes, we assumed that the filing which was submitted last contained the correct information. However if one of the duplicates contained information on a lis pendens filing, we retained that information.

Using this method, we found a total of 214,705 unique loans and 33,859 duplicates in the PFF dataset. From there, we removed records that contained obvious errors (e.g. loans that were originated in the future) and records of 90-day letters that were not mailed in the year 2010. This reduced the PFF dataset to 211,962 clean records.

To ensure comparability across loans, we chose to focus on first-lien mortgages. This reduced the PFF dataset to 186,366 records, but it was a necessary step because a first-lien mortgage is very different from a home equity line of credit (HELOC). The former is frequently taken for the purpose of purchasing a home, while the latter is often used for home improvement.

Our analysis pays particular attention to the 130,912 first-lien mortgages that were originated in the years 2004-2008. Table 1 shows that these five years account for 70 percent of all PFF filings on first-lien mortgages.

We chose to work with the years 2004-2008 because we wanted to compare the PFF data to the data on originations from the Home Mortgage Disclosure Act (HMDA). We chose 2004 as the first year, because the variables available in the pre-2004 HMDA data were quite limited. At the time of this writing, the 2009 HMDA data was available to us, but we chose not to work with it because lending practices changed dramatically after the subprime mortgage crisis crippled the world financial system in late 2008. Loans originated in 2009 were very different from loans originated in previous years, so – for this analysis – we wanted to focus on loans originated in the years leading up to and including the crisis. In a forthcoming analysis, we will compare lending patterns in the periods before and after the crisis and how those differences affect the rate at which borrowers default on their home mortgages.

[3] In cases where the servicer did not include a loan number, we used the property address instead of the loan number.

Table 1: Distribution of Pre-Foreclosure Filings by Year of Origination

	total	percent
1976-1989	2,502	1.3%
1990-1999	13,692	7.3%
2000	2,414	1.3%
2001	4,390	2.4%
2002	7,470	4.0%
2003	16,706	9.0%
2004	18,669	10.0%
2005	28,506	15.3%
2006	35,947	19.3%
2007	31,771	17.0%
2008	16,019	8.6%
2009	6,957	3.7%
2010	1,323	0.7%
total	186,366	100.0%
Data: Full PFF		

3.2 Matching the Pre-Foreclosure Filing Data to the HMDA Originations Data

The HMDA originations data contains the FIPS county code and census tract number of each property. This is a particularly valuable piece of information because census tracts have a small population (between 2,500 and 8,000 people) which is fairly homogeneous in terms of socio-economic characteristics and living conditions (U.S. Census Bureau, 2000).

So our first step in matching the PFF data to the HMDA data was to use the address information to identify the census tract of each property in the PFF dataset. To identify the census tracts, we used Erle's (2005) "Geo-Coder-US-1.00" Perl module in conjunction with the U.S. Census Bureau's (2007) TIGER/Line Files.

After using Erle's Perl module to create a database of New York State addresses from the TIGER/Line Files, we queried the database to obtain the latitudes and longitudes of the property addresses in the PFF dataset. Once we had the coordinates, we compared them to a database of census tract coordinates that we generated from the U.S. Census Bureau's (2005) "Cartographic Boundary Files."

Using this method, we were able to identify the census tracts for 96 percent[4] of the addresses in the PFF database. To avoid losing the information that the other four percent contain, we identified each of the census tracts within the property's five-digit zip code and counted the number of times each census tract corresponded to that zip code. We then randomly assigned the property to one of those census tracts (using the number of occurrences as weights).

[4] 238,830 of the 248,556 (non-unique) addresses

Table 2: Pre-Foreclosure Filings by Loan Amount

	no PFF	received PFF	percent
under 50	4.9%	2.8%	4.8%
50 to 99	16.5%	13.4%	16.3%
100 to 249	36.1%	27.7%	35.4%
250 to 399	25.8%	33.7%	26.4%
400 to 499	8.3%	12.7%	8.6%
500 and up	8.4%	9.7%	8.5%
total	1,544,118	130,722	1,674,840
Data: Combined HMDA-PFF			

Table 3: Pre-Foreclosure Filings by Applicant Income

	no PFF	received PFF	percent
under 40	10.9%	9.9%	10.8%
40 to 59	18.0%	15.6%	17.8%
60 to 79	19.2%	18.3%	19.1%
80 to 99	15.8%	17.3%	15.9%
100 to 119	10.9%	12.9%	11.1%
120 to 159	11.9%	14.0%	12.0%
160 to 199	5.0%	5.4%	5.0%
200 and up	8.4%	6.6%	8.2%
total	1,465,078	123,878	1,588,956
Data: Combined HMDA-PFF			

Once the Census Tracts of each property had been identified and we had purged the duplicates, matching the pre-foreclosure filing data to the HMDA originations data was fairly simple. We divided owner-occupied[5], first-lien mortgages in the HMDA data and first-lien mortgages in the PFF data into buckets by year of origination, census tract and co-applicant status. On average, there were 34 loans in each HMDA bucket and 3 loans in each PFF bucket, so to figure out which HMDA origination corresponded to the pre-foreclosure filing, we compared the loan amounts and chose the closest match.

[5]Mortgage servicers only file pre-foreclosure filing notices when the property is a primary residence, so when matching the PFF data to the HMDA data, we focus on mortgages originated for owner-occupied properties.

Table 4: Pre-Foreclosure Filings by Loan Cost

	no PFF	received PFF	total
non-high cost	92.8%	7.2%	1,364,557
high cost	89.4%	10.6%	310,283
percent	92.2%	7.8%	1,674,840
Data: Combined HMDA-PFF			

Table 5: High Cost Loans by Applicant Income

	non-high cost	high cost	percent
under 40	10.1%	13.9%	10.8%
40 to 59	17.8%	18.0%	17.8%
60 to 79	19.0%	19.4%	19.1%
80 to 99	15.6%	16.9%	15.9%
100 to 119	10.8%	12.1%	11.1%
120 to 159	12.1%	11.9%	12.0%
160 to 199	5.3%	4.1%	5.0%
200 and up	9.3%	3.8%	8.2%
total	1,290,774	298,182	1,588,956
Data: Combined HMDA-PFF			

4 Who Defaults on their Home Mortgage?

4.1 Financial Characteristics

Using the combined HMDA-PFF data, we find that the best predictor that a borrower would default is the amount borrowed. As table 2 shows, 56 percent of the borrowers who received a pre-foreclosure filing took loans in excess of $250,000, whereas only 43 percent of the borrowers who did not default borrowed more than $250,000.

It would be particularly insightful to compare the amounts that borrowers owe to the value of their homes. Unfortunately, HMDA does not provide the loan-to-value ratio or any information on the down payment, so we cannot make such a comparison. Nonetheless, if individuals who borrowed less have a larger equity stake in their homes, then these findings would illustrate the general principle that borrowers who have a larger equity stake in their home are less likely to default and enter the foreclosure process.

Repaying a mortgage also depends on the ability to pay, of course. But it's particularly striking to note that "middle-income" borrowers received pre-foreclosure filings at a higher rate than low and high-income borrowers. Specifically, as table 3 shows, the distribution of income among borrowers who defaulted was

Table 6: High Cost Loans by Loan Amount

	non-high cost	high cost	percent
under 50	4.4%	6.6%	4.8%
50 to 99	15.9%	18.0%	16.3%
100 to 249	37.1%	28.1%	35.4%
250 to 399	25.8%	29.0%	26.4%
400 to 499	8.1%	10.8%	8.6%
500 and up	8.7%	7.6%	8.5%
total	1,364,557	310,283	1,674,840
Data: Combined HMDA-PFF			

Table 7: High Cost Loans by Additional Applicant

	non-high cost	high cost	total
no co-applicant	77.8%	22.2%	952,877
co-applicant	86.3%	13.7%	721,963
percent	81.5%	18.5%	1,674,840
Data: Combined HMDA-PFF			

more heavily weighted in the $80,000 to $199,999 range than the comparable distribution of borrowers who did not default[6].

Why middle-income borrowers default at a higher rate than low-income borrowers is puzzling. The regression models discussed in section 5 suggest however that borrowers with higher incomes are less likely to receive a pre-foreclosure filing after controlling for other factors, such as: loan amount, predicted rate spread, changes in county-level employment and changes in the MSA-level home price index.

Another good predictor of default is the interest cost of the loan. Table 4 shows that borrowers who took "high-cost" loans were more likely to receive a pre-foreclosure filing. When viewed in a risk-premium context, this finding should not be surprising. Borrowers who are more likely to default will have to compensate the lender for the additional risk by paying a higher interest rate.

However, there is also a risk that the additional cost of the loan will make the borrower more likely to default and go into foreclosure. In particular, a borrower's monthly payment is an increasing function of the interest rate, so a higher interest rate reduces a borrower's ability to repay the loan.

Lenders do not set interest rates exogenously however. Since a borrower's income and loan amount affect his/her probability of default, all else equal one would expect lenders to compensate for the addi-

[6]Using a range of $80,000 to $199,999 is an odd way to define "middle income." The range corresponds to the break points in table 3.

Table 8: Pre-Foreclosure Filings by Additional Applicant

	no PFF	received PFF	total
no co-applicant	91.1%	8.9%	952,877
co-applicant	93.6%	6.4%	721,963
percent	92.2%	7.8%	1,674,840
Data: Combined HMDA-PFF			

Table 9: High Cost Loans by Applicant Race

	non-high cost	high cost	total
Asian	89.7%	10.3%	89,998
Black/Afr. Am.	64.9%	35.1%	166,380
White	84.2%	15.8%	1,161,960
not provided	76.8%	23.2%	234,393
percent	81.5%	18.5%	1,674,840
Data: Combined HMDA-PFF			

tional risk by charging a higher interest rate to low-income borrowers and borrowers who take out a larger loan.

In line with this reasoning, we find that low-income borrowers are more likely to receive a high-cost loan than borrowers with higher income. Table 5 shows 80 percent of high-cost loans were originated to borrowers with income below $120,000, whereas only 73 percent of loans that were not high-cost loans were originated to such borrowers.

Surprisingly however, there does not appear to be any systematic relationship between loan amount and the likelihood of the loan being a high-cost loan. Table 6 shows that loan amounts below $100,000 were more likely to be high-cost loans and loan amounts in the $250,000 to $499,999 range were also more likely to be high-cost loans.

It is difficult to understand why small loan amounts (i.e. those under $100,000) were more likely to be high-cost loans and why large loan amounts (i.e. those over $500,000) were less likely to be high-cost loans. Regression analysis does not even help to explain this puzzle. As discussed in section 5, borrowers who took out larger loan amounts tended to receive lower interest rates on their mortgages after controlling for other factors even though the larger loan amounts made them more likely to default.

Another important factor in explaining interest rates is whether there is a co-borrower on the loan or not. As table 7 shows, 22 percent of loans without a co-applicant were high-cost loans, whereas only 14 percent of loans with a co-applicant were high-cost loans. This may be attributable to the fact that a second borrower is a (potential) second source of income, which helps to mitigate the risk that the loan

Table 10: High Cost Loans by Applicant Ethnicity

	non-high cost	high cost	total
Hispanic/Latino	71.9%	28.1%	134,937
Not Hispanic/Latino	82.8%	17.2%	1,263,971
not provided	77.5%	22.5%	232,693
percent	81.5%	18.5%	1,674,840
Data: Combined HMDA-PFF			

Table 11: Pre-Foreclosure Filings by Applicant Race

	no PFF	received PFF	total
Asian	92.8%	7.2%	89,998
Black/Afr. Am.	88.0%	12.0%	166,380
White	92.8%	7.2%	1,161,960
not provided	91.7%	8.3%	234,393
percent	92.2%	7.8%	1,674,840
Data: Combined HMDA-PFF			

will go into default. As table 8 shows, 9 percent of loans without a co-borrower received a pre-foreclosure filing, whereas only 6 percent of loans with a co-borrower received a pre-foreclosure filing.

4.2 Race and Ethnicity

In section 2, we reviewed evidence of racial and ethnic discrimination in lending practices. The HMDA data captures one form of such discrimination – the difference in the rate spread between loans originated to minorities and loans originated to whites. As tables 9 and 10 show, blacks and Latinos received a disproportionately high share of high-cost loans. Asians, by contrast, received a disproportionately low share.

Tables 11 and 12 show that blacks and Latinos also received a disproportionately high share of pre-foreclosure filings, so one also has to wonder if racial and ethnic discrimination in lending practices contributed to the disproportionately high share of defaults among blacks and Latinos.

One way to address this question is to ask if fundamental differences between minorities and non-minorities justify the difference in rate spreads. If so, then the next question to ask is if those fundamental differences could have caused blacks and Latinos to default at disproportionately higher rates.

The first fundamental factor that we'll consider is income. If minority borrowers tended to have lower

Table 12: Pre-Foreclosure Filings by Applicant Ethnicity

	no PFF	received PFF	total
Hispanic/Latino	89.0%	11.0%	134,937
Not Hispanic/Latino	92.4%	7.6%	1,263,971
not provided	92.0%	8.0%	232,693
total	92.2%	7.8%	1,674,840
Data: Combined HMDA-PFF			

Table 13: Applicant Income by Applicant Race

	Asian	Black/Afr. Am.	White	not provided	percent
under 40	4.0%	8.0%	12.2%	8.7%	10.8%
40 to 59	11.7%	16.5%	18.9%	16.1%	17.8%
60 to 79	16.3%	23.0%	18.7%	19.4%	19.1%
80 to 99	17.3%	20.1%	15.1%	16.0%	15.9%
100 to 119	14.4%	13.6%	10.4%	11.0%	11.1%
120 to 159	17.6%	12.4%	11.5%	12.5%	12.0%
160 to 199	8.3%	3.7%	4.9%	5.5%	5.0%
200 and up	10.5%	2.7%	8.4%	10.8%	8.2%
total	85,965	156,030	1,105,913	220,741	1,588,956
Data: Combined HMDA-PFF					

income than their non-minority counterparts, then one could justify the difference in rate spreads on the basis of income.

Such a hypothesis only finds partial support in the data. Table 13 shows that 26 percent of Asian borrowers and 18 percent of white borrowers had income over $140,000, while only 11 percent of black borrowers did. The distribution of income by ethnicity shows a similar pattern. As table 14 shows, 18 percent of non-Latino borrowers had income over $140,000, while only 14 percent of Latinos did.

The fact that there is more weight in the upper region of the distribution of income among non-minority borrowers than there is in the distribution of income among non-minority borrowers lends some support to the hypothesis that differences in income help explain why blacks and Latinos received a disproportionate share of high-cost loans.

However, the lower region of the income distributions refutes the hypothesis. It appears to have been easier for low-income whites and non-Latinos to obtain a mortgage. Specifically, 31 percent of white borrowers had income below $60,000, while only 25 percent of black borrowers did. Similarly, 30 percent of non-Latinos had income below $60,000, while only 19 percent of Latinos did.

Table 14: Applicant Income by Applicant Ethnicity

	Hispanic/Latino	Not Hispanic/Latino	not provided	percent
under 40	5.8%	11.6%	8.9%	10.8%
40 to 59	12.9%	18.5%	16.3%	17.8%
60 to 79	20.6%	18.9%	19.2%	19.1%
80 to 99	21.4%	15.4%	15.8%	15.9%
100 to 119	15.9%	10.6%	10.9%	11.1%
120 to 159	14.8%	11.7%	12.4%	12.0%
160 to 199	4.8%	5.0%	5.5%	5.0%
200 and up	3.8%	8.2%	11.0%	8.2%
total	125,440	1,203,686	219,669	1,588,956
Data: Combined HMDA-PFF				

Table 15: Loan Amount by Applicant Race

	Asian	Black/Afr. Am.	White	not provided	percent
under 50	1.0%	3.2%	5.7%	2.7%	4.8%
50 to 99	6.3%	8.4%	19.1%	12.1%	16.3%
100 to 249	26.3%	28.3%	37.2%	35.2%	35.4%
250 to 399	33.3%	40.5%	23.0%	29.9%	26.4%
400 to 499	18.0%	12.8%	7.1%	9.4%	8.6%
500 and up	15.1%	6.8%	7.8%	10.7%	8.5%
total	89,998	166,380	1,161,960	234,393	1,674,840
Data: Combined HMDA-PFF					

Consequently, it would be hard to justify the disproportionate share of high-cost loans that blacks and Latinos received on the basis of income differentials.

A second fundamental factor to consider is the amount of the original loan. Differences in loan amounts help explain why blacks and Latinos received a disproportionate share of pre-foreclosure filings, but they do not necessarily explain why they received a disproportionate share of high-cost loans.

Specifically, minorities tended to borrow much more than their non-minority counterparts. According to table 15 shows, 53 percent of white borrowers borrowed less than $200,000 whereas only 28 percent of blacks did. Interestingly however, Asians appear to have borrowed even more than blacks (only 25 percent borrowed less than $200,000), but had the lowest rate of high-cost loans. Turning to ethnicity, table 16 shows that 50 percent of non-Latinos borrowed less than $200,000, whereas 25 percent of Latinos borrowed less than that amount.

The finding that blacks and Latinos tended to borrow more helps explain why they received a disproportionately high share of pre-foreclosure filings, but it does not explain why they took high-cost loans

Table 16: Loan Amount by Applicant Ethnicity

	Hispanic/Latino	Not Hispanic/Latino	not provided	percent
under 50	2.1%	5.4%	2.9%	4.8%
50 to 99	7.1%	17.8%	12.7%	16.3%
100 to 249	26.4%	36.1%	35.6%	35.4%
250 to 399	41.7%	24.4%	29.0%	26.4%
400 to 499	13.6%	8.1%	9.0%	8.6%
500 and up	9.2%	8.1%	10.7%	8.5%
total	134,937	1,263,971	232,693	1,674,840
Data: Combined HMDA-PFF				

at a higher rate than their white, Asian and non-Latino counterparts. Asians borrowed more, but took fewer high-cost loans. Moreover, as mentioned previously, the regression analysis in section 5 also refutes the hypothesis that borrowers who took out larger loan amounts would receive lower interest rates. The opposite is true. All else equal, the rate spreads on larger loans tend to be lower.

5 Econometric Models of Rate Spreads and Defaults

Section 4 describes several questions raised by the PFF data and the combined HMDA-PFF dataset. The most striking questions are why blacks and Latinos were more likely to take high-cost loans and why they are more likely to default on their mortgages. But there were other questions too. One is the lack of a clear relationship between the amount of the original loan and the whether the loan was a high-cost loan or not. Another was why middle-income borrowers were more likely to default than both low-income and high-income borrowers.

In an attempt to resolve some of these puzzles, this section presents a very basic regression analysis, so that we can examine the effect of one variable while holding others constant. The analysis presented here makes no effort to place the variables in a theoretical framework. Nor does it make much effort to check for robustness across specifications. Such work is left to future research.

The regression analysis presented here simply attempts to use the available variables to predict the rate spread on a borrower's loan. The predicted rate spread is then used as an instrument in a second-stage probit regression to estimate a borrower's probability of defaulting on the loan.

Table 17: Two-Stage: Tobit predicts Rate Spread, then Probit predicts PFF

	Model #1				Model #2			
	Tobit		probit		Tobit		probit	
Intercept	−0.0513	***	−2.1133	***	0.0037		−2.1071	***
	(0.0004)		(0.1183)		(0.0054)		(0.1715)	
Pred. Rate Spread			0.4093	.			0.3302	
			(0.2434)				(0.3173)	
ln(Loan Amount)	−0.0005	***	0.2511	***	−0.0005	***	0.2486	***
	(0.0001)		(0.0252)		(0.0001)		(0.0366)	
ln(App. Income)	−0.0014	***	−0.2067	***	−0.0009	***	−0.2054	***
	(0.0001)		(0.0251)		(0.0001)		(0.0365)	
Co-Applicant	−0.0053	***	−0.1044	***	−0.0049	***	−0.1059	**
	(0.0001)		(0.0243)		(0.0001)		(0.0352)	
Conv'l Loan	0.0156	***			0.0158	***		
	(0.0002)				(0.0002)			
Home Purchase	0.0114	***			0.0112	***		
	(0.0001)				(0.0001)			
Home Improve.	0.0075	***			0.0073	***		
	(0.0001)				(0.0001)			
Hispanic/Latino	0.0092	***	0.1705	***	0.0064	***	0.1702	**
	(0.0001)		(0.0424)		(0.0001)		(0.0616)	
Asian	−0.0017	***	−0.0447		−0.0034	***	−0.0456	
	(0.0002)		(0.0510)		(0.0002)		(0.0742)	
Black/Afr. Am.	0.0136	***	0.2381	***	0.0086	***	0.2396	***
	(0.0001)		(0.0395)		(0.0001)		(0.0575)	
Race not provided	0.0060	***	0.0662	*	0.0047	***	0.0640	
	(0.0001)		(0.0334)		(0.0001)		(0.0485)	
Female	0.0019	***	−0.0174		0.0018	***	−0.0180	
	(0.0001)		(0.0249)		(0.0001)		(0.0363)	
Δ ln(County Emp.)			−1.8524	**			−1.9836	*
			(0.5722)				(0.8206)	
Δ ln(House Price Idx.)			−0.3514	.			−0.3530	
			(0.1844)				(0.2678)	
Minority Pop. Pct.					0.0001	***		
					(0.0000)			
ln(HUD Median					−0.0059	***		
Family Income)					(0.0005)			

Continued on the next page.

Table 17 (continued)

	Model #1				Model #2			
	Tobit		probit		Tobit		probit	
Purch. Type = 5	0.0288	***			0.0282	***		
	(0.0001)				(0.0001)			
Purch. Type = 6	0.0114	***			0.0112	***		
	(0.0001)				(0.0001)			
Purch. Type = 7	0.0186	***			0.0183	***		
	(0.0001)				(0.0001)			
Purch. Type = 8	0.0030	***			0.0030	***		
	(0.0001)				(0.0001)			
Purch. Type = 9	0.0196	***			0.0192	***		
	(0.0001)				(0.0001)			
Capital	0.0058	***			0.0132	***		
	(0.0001)				(0.0002)			
Central	0.0065	***			0.0134	***		
	(0.0002)				(0.0002)			
Finger Lakes	0.0058	***			0.0126	***		
	(0.0001)				(0.0002)			
Long Island	0.0012	***			0.0083	***		
	(0.0001)				(0.0002)			
Mid-Hudson	0.0004	***			0.0058	***		
	(0.0001)				(0.0001)			
Mohawk Valley	0.0116	***			0.0182	***		
	(0.0002)				(0.0002)			
North Country	0.0119	***			0.0180	***		
	(0.0002)				(0.0003)			
Southern	0.0099	***			0.0165	***		
	(0.0002)				(0.0002)			
Western	0.0073	***			0.0140	***		
	(0.0001)				(0.0002)			
New York County	−0.0233	***			−0.0206	***		
	(0.0004)				(0.0004)			
orig. 2005	0.0110	***	0.1604	***	0.0111	***	0.1589	**
	(0.0001)		(0.0402)		(0.0001)		(0.0583)	
orig. 2006	0.0146	***	0.3100	***	0.0147	***	0.3096	***
	(0.0001)		(0.0498)		(0.0001)		(0.0723)	
orig. 2007	0.0096	***	0.3678	***	0.0099	***	0.3642	***
	(0.0001)		(0.0542)		(0.0001)		(0.0785)	
orig. 2008	0.0041	***	0.2130	***	0.0049	***	0.2098	**
	(0.0001)		(0.0546)		(0.0001)		(0.0790)	
AIC	−561,338		827,003		−572,134		826,728	
*** $p < 0.001$, ** $p < 0.010$, * $p < 0.050$, . $p < 0.100$								
Standard errors in parenthesis.								
Data: Combined HMDA-PFF								

Table 18: Distribution of Interest Rates in Pre-Foreclosure Filing Data

	total	percent
under 4.000	11,133	6.0%
4.000 to 5.999	49,876	26.8%
6.000 to 7.999	94,870	50.9%
8.000 to 9.999	21,643	11.6%
10.000 to 11.999	7,060	3.8%
12.000 to 13.999	1,430	0.8%
14.000 and up	354	0.2%
total	186,366	100.0%
Data: Full PFF		

Table 19: Pre-Foreclosure Filings by Predicted Interest Rate (Tobit Model #1)

	no PFF	received PFF	percent
under 4.000	18.9%	13.2%	18.4%
4.000 to 5.999	49.2%	47.0%	49.0%
6.000 to 7.999	26.0%	30.4%	26.4%
8.000 to 9.999	5.8%	9.1%	6.0%
10.000 to 11.999	0.1%	0.3%	0.1%
total	1,435,566	122,402	1,557,968
Data: Combined HMDA-PFF			

5.1 Econometric Methods

One problem confronting any econometric analysis of the HMDA data is how to work with the rate spread. The HMDA data only provides a value of the rate spread when the difference between the interest rate on the mortgage and the yield on the comparable U.S. Treasury exceeds three percentage points[7]. Consequently, when addressing the question of why black and Latino borrowers were more likely to take out a high-cost loan, we have to find a way to work with the rate spread.

The simplest method is to reduce the rate spread to a binary variable (i.e. one if high-cost, zero otherwise) and employ a probit or logit model to estimate the probability that a borrower took a high-cost loan. The trouble with such a strategy is that is purges valuable information on the magnitude of the differences in rate spread among borrowers.

The alternative is to employ a Tobit model to obtain an estimate of the rate spread itself. The trouble

[7]More precisely, the HMDA data provides a value for the rate spread of a first-lien mortgage when it exceeds three percentage points. For other lien statuses, the HMDA data provides a value for the rate spread when it exceeds five percentage points. Our analysis focuses exclusively on first-lien mortgages.

Table 20: Pre-Foreclosure Filings by Predicted Interest Rate (Tobit Model #2)

	no PFF	received PFF	percent
under 4.000	19.4%	13.7%	19.0%
4.000 to 5.999	48.4%	45.6%	48.2%
6.000 to 7.999	26.0%	30.9%	26.4%
8.000 to 9.999	6.0%	9.6%	6.3%
10.000 to 11.999	0.1%	0.3%	0.2%
total	1,435,566	122,402	1,557,968
Data: Combined HMDA-PFF			

with this strategy is that 81 percent of the loans in the combined HMDA-PFF dataset are not high-cost loans, so no value of the rate spread is reported for these loans. Therefore, instead of using the Tobit model to estimate the tail of the distribution, the Tobit model has to estimate 81 percent of the distribution.

We chose to use the Tobit model however because it provides an estimate of the rate spread which can be used as an instrument in a second-stage regression on the probability of defaulting on the home mortgage. One must use an instrument for the rate spread in the second-stage to overcome the endogeneity problem that arises when lenders charge higher interest rates to borrowers who are more likely to default.

To obtain efficient estimates of the parameters in the second-stage probability model, we used an algorithm that Adkins (2009) developed to implement Amemiya's Generalized Least Squares (AGLS). Adkins (2008) shows that the AGLS estimator yields consistent estimates of the parameters' standard errors and can be used to test the statistical significance of the parameters.

The AGLS algorithm requires estimates of the residuals from the first-stage regression, but – because the rate spread is censored at three percentage points – we could not use response residuals as we would if the first-stage regression were a standard OLS regression model. Therneau and Lumley's (2009) "survival" package for R (R Development Core Team, 2010) provides a viable alternative however. As its "survreg" function iteratively maximizes the log-likelihood function, it predicts the value of the dependent variable and calculates a correction term, called the "working residual" (Therneau, 1999), which we use in place of the response residual.

5.2 Discussion of the Regression Results

As shown in table 17, the rate spreads on owner-occupied, first-lien mortgages originated to blacks and Latinos were higher than those originated to their white and non-Latino counterparts and the differences were statistically significant, even after controlling for other variables such as income, loan amount, whether there was a co-borrower on the loan, the purpose of the loan and region of the state and year of origination.

Importantly, the racial and ethnic disparities in interest rates were large. The coefficient estimates in model #1 suggest that the interest rate on a loan originated to a black borrower was 1.36 percentage points higher than a the interest rate originated to an equivalent white borrower. Model #2 suggests a slightly smaller difference: 0.86 percentage points. Turning to Latinos, the coefficient estimates in model #1 suggest that Latinos paid 0.92 percentage points more than an equivalent non-Latino borrower, while model #2 puts the gap at 0.64 percentage points.

While this is deeply disturbing, the HMDA data omits many important variables (such as the borrower's credit score and the loan-to-value ratio), so we are reluctant to conclude that this is evidence of discrimination.

With one exception, the signs of the other coefficients in the model are not surprising. The coefficient on loan amount is the exception. It seems odd to us that borrowers who took out larger loans would pay a lower interest rate. In the case of the HMDA data however, a large loan amount may be acting as a proxy for variables that we do not observe and thus indicate that the borrower is more creditworthy.

Before accepting our findings at face value however, one must note an important limitation of using the Tobit model to predict the rate spread: the estimates are far from perfect. By adding the average yield on a 30-year U.S. Treasury to the predicted rate spread, we can compare the Tobit models' predicted interest rates to the ones in the pre-foreclosure filing data. As tables 18, 19 and 20 show, the predicted interest rates do not have as much weight in the upper region as the interest rates in the Full PFF dataset. We believe that the predicted rate spread is correlated with the unobserved true values of the rate spread, but there is no way to check the validity of this assumption.

Turning to the second-stage model of the probability that a borrower will default, we find that the coefficient on the predicted rate spread is positive (suggesting that borrowers with higher rate spreads were more likely to default), but is only statistically significant at the 10 percent level in model #1 and is not statistically significant at all in model #2.

Both models suggest that black race and Latino ethnicity are positively correlated with the probability of default after controlling for other factors, such as income, loan amount and whether there is a co-applicant on the loan. We do not believe however that this reflects personal characteristics. Instead, we believe that the limitations of the HMDA data are causing black race and Latino ethnicity to act as a proxy for a missing variable. Given our review of the evidence of discrimination in lending practices in section 2, one possibility is that black race and Latino ethnicity are acting as a proxy for a form of discrimination that we cannot measure with the HMDA data.

Both of the models also predict that borrowers who took out larger loans were more likely to default after controlling for other factors. This finding coupled with table 2's finding that borrowers who took loans in excess of $250,000 were much more likely to default than those who borrowed less leads us to conclude that large loan amounts are the best predictor of default.

The signs of the coefficients on other variables were in line with expectations, but it is important to note that the coefficient on applicant income was negative and statistically significant in both models. This finding helps us explain one of the puzzles that we observed in section 4: the puzzle that middle-income

borrowers were more likely to default (as shown in table 3).

We could have used a quadratic term in the regression model to reproduce the result in table 3, but given the possibility that income is correlated with some of the other explanatory variables, we were reluctant to over-fit the model. Testing the statistical significance of the coefficient on a quadratic term is left to future research.

It is also interesting to note that the coefficient on the percentage change in the home price index is only statistically significant at the 10 percent level in model #1 and is not statistically significant at all in model #2. By contrast, the coefficient on the percentage change in county-level employment is statistically significant at the 5 percent level in both models. This suggests that job losses are a better predictor of default than decreasing home values.

6 Conclusion

Matching the New York State pre-foreclosure filing data to the HMDA originations data reveals the same racial and ethnic disparities in lending practices that other studies have found, but that finding provides very little insight into how one can reduce the rate at which borrowers default on their mortgages.

Given our finding that large loan original amounts are the best predictor that a borrower would default on his/her home mortgage, one could conclude that reducing the principal balances on home mortgages would substantially reduce the rate at which borrowers default.

Reducing principal balances may be impractical, however. In cases where borrowers have negative equity, this would require lenders to absorb potentially very large losses on their portfolio of mortgages. Secondly, an across-the-board reduction in principal balance would benefit a large number of borrowers who otherwise would not default on their mortgages.

Moreover, it is not the default per se that imposes a financial burden on lenders. It is the default that progresses to foreclosure that reduces the value of a lender's portfolio of home mortgages (when the lender is unable to recover the principal balance from the proceeds of the foreclosure sale).

We analyze the issues associated with foreclosure prevention in a second paper (Doviak and MacDonald, 2011). In that paper, we find that the original loan amount is positively correlated with the probability of progressing from default to foreclosure. Consequently, reducing principal balances is might have the desired effect of reducing losses in the mortgage industry if the modifications were well-structured, so that the balance-sheet effect of the lower probability of progressing to foreclosure offsets the losses that the lender would suffer by taking the loan to foreclosure.

Assuming that such a structure could be found, it may depend on information that the pre-foreclosure filing data does not contain, such as the borrower's income or the purpose of the loan. The HMDA data does contain this information however, so in future work, we plan to incorporate the information from

the HMDA dataset into the Short PFF dataset to see how those factors affect a borrower's probability of progressing from default to foreclosure and explore other options that may help the industry reduce its losses. In that analysis, we will also attempt to quantify the savings that the industry would achieve from such modifications.

7 Acknowledgements

We would like to thank the New York State Banking Department for making the Pre-Foreclosure Filing data available to us and for supporting our research. The views expressed in this paper are our own opinions and do not necessarily reflect the opinions of the New York State Banking Department.

References

L. C. Adkins. "Small Sample Performance of Instrumental Variables Probit Estimators: A Monte Carlo Investigation". *Joint Statistical Meetings Proceedings, Business and Economics Statistics Section*, 4 Aug. 2008.

L. C. Adkins. "An Instrumental Variables Probit Estimator using gretl". *Econometrics with gretl, Proceedings of the gretl Conference 2009*, pages 59–74, 28-29 May 2009.

D. G. Bocian, K. S. Ernst, and W. Li. "Unfair Lending: The Effect of Race and Ethnicity on the Price of Subprime Mortgages", 31 May 2006. URL http://www.responsiblelending.org/mortgage-lending/research-analysis/rr011-Unfair_Lending-0506.pdf.

C. Bromley, J. Campen, S. Nafici, A. Rust, G. Smith, K. Stein, and B. van Kerkhove. "Paying More for the American Dream: The Subprime Shakeout and Its Impact on Lower-Income and Minority Communities", Mar. 2008. URL http://www.policyarchive.org/handle/10207/19021.

M. Doms, F. Furlong, and J. Krainer. "Subprime Mortgage Delinquency Rates". Working Paper 2007-33, Nov. 2007. URL http://www.frbsf.org/publications/economics/papers/2007/wp07-33bk.pdf.

E. Doviak and S. MacDonald. "Who Enters the Foreclosure Process?". 24 Sept. 2011. URL http://www.doviak.net/foreclosure/foreclosure.htm.

S. Erle. "Geo-Coder-US-1.00: Geocode (estimate latitude and longitude for) any US address". Comprehensive Perl Archive Network, 17 May 2005. URL http://search.cpan.org/~sderle/Geo-Coder-US-1.00/.

K. S. Gerardi and P. S. Willen. "Subprime Mortgages, Foreclosures, and Urban Neighborhoods". Public Policy Discussion Papers, no. 08-6, 22 Dec. 2008. URL http://www.bos.frb.org/economic/ppdp/2008/ppdp0806.pdf.

E. Laderman. "Subprime Mortgage Lending and the Capital Markets". *FRBSF Economic Letter*, 2001(38), 28 Dec. 2001. URL http://www.frbsf.org/publications/economics/letter/2001/el2001-38.pdf.

D. P. Morgan, B. Iverson, and M. Botsch. "Subprime Foreclosures and the 2005 Bankruptcy Reform". *FRBNY Economic Policy Review*, 2011. Forthcoming.

New York State Banking Department. "2009 Mortgage Foreclosure Law - Overview", 2009. URL http://www.banking.state.ny.us/mfl2009.htm.

R Development Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2010. URL http://www.R-project.org. ISBN 3-900051-07-0.

J. S. Rugh and D. S. Massey. "Racial Segregation and the American Foreclosure Crisis". *American Sociological Review*, 75(5):629–651, 2010.

G. D. Squires, D. S. Hyra, and R. N. Renner. "Segregation and the Subprime Lending Crisis". EPI Briefing Paper, 4 Nov. 2009. URL http://www.epi.org/page/-/pdf/110409-briefingpaper244.pdf.

T. Therneau and T. Lumley. *survival: Survival analysis, including penalised likelihood.*, 2009. URL http://CRAN.R-project.org/package=survival. R package version 2.35-8.

T. M. Therneau. "A Package for Survival Analysis in S", 27 Jan. 1999.

U.S. Census Bureau. "Census Tracts and Block Numbering Areas", 19 Apr. 2000. URL http://www.census.gov/geo/www/cen_tract.html.

U.S. Census Bureau. "Cartographic Boundary Files", 27 June 2005. URL http://www.census.gov/geo/www/cob/tr2000.html.

U.S. Census Bureau. "2006 Second Edition TIGER/Line Files", 5 Mar. 2007. URL http://www.census.gov/geo/www/tiger/tiger2006se/tgr2006se.html.

U.S. Dept. of Housing and Urban Development. "Unequal Burden: Income and Racial Disparities in Subprime Lending", May 2000a. URL http://archives.hud.gov/reports/subprime/subprime.cfm.

U.S. Dept. of Housing and Urban Development. "Unequal Burden in New York: Income and Racial Disparities in Subprime Lending", May 2000b. URL http://www.huduser.org/Publications/pdf/newyork.pdf.

A Additional Tables

Table 21: High Cost Loans by Year of Origination

	non-high cost	high cost	percent
2004	28.1%	17.9%	26.2%
2005	22.7%	30.2%	24.1%
2006	18.8%	29.3%	20.7%
2007	17.3%	15.9%	17.1%
2008	13.1%	6.7%	11.9%
total	1,364,557	310,283	1,674,840
Data: Combined HMDA-PFF			

Table 22: Pre-Foreclosure Filings by Year of Origination

	no PFF	received PFF	percent
2004	27.2%	14.3%	26.2%
2005	24.3%	21.8%	24.1%
2006	20.2%	27.5%	20.7%
2007	16.4%	24.3%	17.1%
2008	11.8%	12.2%	11.9%
total	1,544,118	130,722	1,674,840
Data: Combined HMDA-PFF			

Table 23: Pre-Foreclosure Filings by Region

	total	percent
Capital	11,700	6.3%
Central	7,259	3.9%
Finger Lakes	12,641	6.8%
Long Island	46,658	25.0%
Mid-Hudson	28,487	15.3%
Mohawk Valley	5,106	2.7%
New York City	52,809	28.3%
North County	2,952	1.6%
Southern Region	5,376	2.9%
Western Region	13,378	7.2%
total	186,366	100.0%
Data: Full PFF		

Table 24: Distribution of Loan Amounts in the PFF Data

	total	percent
under 50	11,629	6.2%
50 to 99	34,020	18.3%
100 to 149	22,091	11.9%
150 to 199	17,065	9.2%
200 to 249	16,426	8.8%
250 to 299	18,079	9.7%
300 to 349	18,938	10.2%
350 to 399	15,640	8.4%
400 to 449	11,524	6.2%
450 to 499	6,553	3.5%
500 and up	14,401	7.7%
total	186,366	100.0%
Data: Full PFF		