Title
Essays on Information and Beliefs in Credit Markets

Author
Botsch, Matthew

Publication Date
2014

Peer reviewed|Thesis/dissertation

Essays on Information and Beliefs in Credit Markets

by

Matthew Jason Botsch

A dissertation submitted in partial satisfaction of the
requirements for the degree of
Doctor of Philosophy

in

Economics

in the

Graduate Division
of the
University of California, Berkeley

Committee in charge:

Professor Ulrike Malmendier, Chair
Professor Ross Levine
Professor David Romer

Spring 2014

Essays on Information and Beliefs in Credit Markets

Abstract

Essays on Information and Beliefs in Credit Markets

by

Matthew Jason Botsch
Doctor of Philosophy in Economics

University of California, Berkeley

Professor Ulrike Malmendier, Chair

This dissertation is a collection of three essays in financial economics, specifically focused on the role of information and beliefs in credit markets. The first chapter establishes that private bank information about customers in primary lending markets exists. The second chapter shows that private information hinders banks' capacities to sell loans on secondary markets, unless the purchaser believes that the bank has committed to remain uninformed. The third chapter explores the welfare consequences of incorrect borrower beliefs about the economic environment on financial product choice.

In the first chapter, my co-author and I hypothesize that while lending to a firm, a bank receives signals that allow it to learn and better understand the firm's fundamentals; and that this learning is private; that is, it is information that is not fully reflected in publicly-observable variables. We test this hypothesis using data from the syndicated loan market between 1987 and 2003. We construct a variable that proxies for firm quality and is unobservable by the bank, so it cannot be priced when the firm enters our sample. We show that the loading on this factor in the pricing equation increases with relationship time, hinting that banks are able to learn about firm quality when they are in an established relationship with the firm.

In the second chapter, I present new evidence that lemon problems hinder trade on secondary mortgage markets. Using the geographic distance from lenders to borrowers as a proxy for the absence of private bank information, I document a systematic positive link between distance and the mortgage sale rate. Mortgage sale rates are higher when the originating lender is less likely to be informed about the borrower. I further show that the private mortgage sale rate locally depends on lender-borrower distance only above the conforming loan limit, in the illiquid jumbo market where the GSEs are barred from purchasing mortgages. This is consistent with the familiar tradeoff between market liquidity and seller incentives to acquire information.

In the third chapter, I investigate how borrowers' incorrect beliefs about future inflation might bias their choice between fixed-rate and adjustable-rate mortgages.

Borrowers who have experienced recent periods of greater inflation pay more for fixed-rate mortgage contracts and pay more in interest, at least over the first six yeras of the mortgage's life. That is, incorrect beliefs about future inflation are welfare-reducing both ex ante and ex post.

To Emil J. Botsch, Marjorie A. Sears, and Robert F. Sears. May the best of them live on in those who remember them.

Contents

Preface

This dissertation is a collection of three essays in financial economics, specifically focused on the role of information and beliefs in credit markets. The unifying theme is how information, and perceptions about everyone else's information, affects credit market participants' behavior. The fundamental problem in lending is the presence of asymmetric information between borrowers and lenders. The first chapter is an empirical exercise that explores how banks are able to resolve this asymmetry and learn about their customers over time in the context of a repeat relationship. However, more bank information about customers is not always privately optimal, particularly if the bank wishes to sell a loan to a less-informed outside counterparty. The second chapter shows how asymmetric infomation between buyers and sellers can hinder trade in secondary markets and incentivize banks to commit "not to learn" about their customers. The third chapter changes tack and focuses on errors in aggregating information into expectations. Incorrect borrower beliefs about future inflation contribute to incorrect decisions between fixed-rate and adjustable-rate mortgages which are welfare-reducing.

The first chapter in this dissertation, "Relationship Lending: Do Banks Learn?" is joint work with Victoria Vanasco. The setting of this chapter is the syndicated loan market, in which large, usually publicly-held firms take out loans from consortiums of banks. We obtain data on the identity of participants and terms of lending for a large proportion of the syndicated loans originated between 1989 and 2003, and link this information with firm financial characteristics from Compustat. Our research question is one of existence: do banks acquire information about the customers with whom they repeatedly interact, and is this information private to the bank or is it revealed to other market participants? Our methodology uses a proxy for borrower quality based on each borrowing firm's cumulative abnormal stock return during the Lehman crisis. We pick this date because it is five years after the last loan in our sample, guaranteeing that the bank cannot have observed the proxy in real time, and because this period is likely to have been particularly revealing in separating the performance of "good" or resilient firms from "bad" or less resilient ones. We find that lending banks increasingly price on this proxy over the course of a lending relationship, indicating that lenders do learn about borrowers over time. Moreover, some of the learning is private: when firms switch banks, the new bank does not price on the firm proxy as much as the original bank. Our estimates indicate an approximate 50-50 split between public and private learning. Some information, but not all, is revealed to outside lenders by the fact that the original bank continues to do business with the borrower.

The second chapter, "Distance, Asymmetric Information, and Mortgage Securitization," focuses on the private information of home buyers vis-à-vis originating banks and of banks vis-à-vis purchasing institutional investors. The theoretical framework is one where banks can exert costly but unobservable effort to audit the loan appli-

cants. If bank learning about the borrower improves loan performance, then auditing is advantageous to any potential purchaser of the loan. On the other hand, if banks can use the private information they acquire while auditing to screen between good type and bad type borrowers, then auditing creates a lemon problem in secondary markets and hinders trade. The equilibrium amount of trade hinges on purchasers forming a correct set of beliefs about how much private information banks are likely to acquire. I test the magnitude of these two effects by using the geographic distance between the bank and U.S. mortgage borrowers over 1990 to 2000 as a proxy for the cost of information acquisition. I find that the lemon effect dominates the quality effect: secondary market sale rates are higher for the loans of faraway borrowers, who are more costly to audit, and lower for the loans of nearby borrowers, who are less costly to audit. This suggests that secondary mortgage markets operate in a "low information equilibrium," wherein banks are incentivized to make loans to customers about whom they credibly do not know very much.

The third and final chapter, "The Welfare Consequences of Experienced Inflation on Residential Mortgage Choice," focuses on the borrower side of the mortgage market. I investigate how borrowers' incorrect beliefs about future inflation might bias their choice of mortgage products. A literature in psychology and economics suggests that individuals overweight recent experiences relative to the optimal Bayesian scheme. For example, young borrowers coming of age during the 1970s have recently experienced a period of high inflation, and they do not have personal memory of earlier periods of lower inflation. If this high experienced inflation translates into a forecast of high future inflation, then these borrowers will demand greater insurance against increases in nominal interest rates. I estimate that every additional percentage point of experienced inflation increases a borrower's willingness to pay for a fixed rate mortgage by 6 to 8 basis points of the FRM contract rate, as compared to an adjustable rate mortgage. Since fully rational individuals should place a weight of zero on the inflation they have personally experienced, these biased expectations have a major impact on the product mix of FRMs versus ARMs. Moreover, this bias is welfare-reducing ex post. I run a simple simulation exercise to calculate that borrowers are overpaying by an average of $220 per year for the embedded inflation insurance of the FRM.

Acknowledgments

The papers in this dissertation would have been much weaker were it not for the amazing support of many other individuals. Herein I attempt to acknowledge many of these people; my apologies to those of you whom I have forgotten.

(My mentioning someone should not be construed to mean that they agree with the conclusions of these papers. Moreover, the usual disclaimer in economics, that any errors in this dissertation are entirely my responsibility, applies here in triplet.)

Thanks first of all to my dissertation adviser, Professor Ulrike Malmendier. Her keen sense of what research questions are interesting kept me going back to the drawing board, perhaps more times than I would have liked, but I believe the end result was much improved because of it. Her feedback and encouragement were invaluable, particularly during the last year of my Ph.D. when I was "on the market."

Thanks to the other two members of my dissertation committee, Professors Ross Levine and David Romer. Their willingness to give comments when I had written a draft and to provide advice when I was in need was both incredibly generous and immensely helpful. They continually pushed me to push my papers in better directions.

Thanks also to Professor James Wilcox, who served on my orals committee and was always willing to listen when I had a new research idea. His enthusiasm about research is contagious.

Bart Hobijn provided great encouragement while I served as a section leader for one of his classes. He was kind enough to invite me to present Chapter 2 at the San Francisco Fed, and he provided very useful stylistic comments on the paper and slides.

Thanks to the current and former staff members of the Economics Department at Berkeley. Particular thanks to the Lead Graduate Adviser, Patrick Allen, who helped me navigate the administrative bureaucracy from day one, and who displayed endless patience on the occasions when I missed a deadline for filing paperwork. Thanks also to Rowilma Balza del Castillo, Vicky Lee, Alex Mastrangeli, Heather Reed, Emil Schissel, Joe Sibol, Phil Walz, and to June Wong at Haas. The Department was a pleasant place to spend these many years in large part because of their tireless work behind the scenes.

Thanks to my officemates, classmates, housemates, and friends (these are non-disjoint sets).

In addition to the above-mentioned, certain people made particular contributions to some of the chapters.

Chapter 1: The idea for this paper originated during a conversation with Michel Serafinelli. The authors would additionally like to thank Vladimir Asriyan, Gabriel Chodorow-Reich, Chad Kendall, Xiaoyu Xia, and seminar participants at Berkeley economics, Berkeley Haas, and EconCon 2013 for helpful conversations and comments.

Chapter 2: Thanks for helpful discussions and comments from Benjamin Iverson, Dwight Jaffee, Amir Kermani, Victoria Vanasco, James Vickery, and seminar participants at Berkeley's Haas School of Business and the Federal Reserve Bank of San Francisco.

Thanks finally to my family. Though I moved thousands of miles to the opposite side of the country for graduate school, you were always there for me, and you never doubted me even when I felt nothing but doubts.

Chapter 1

Relationship Lending: Do Banks Learn?[1]

[1]This chapter is joint work with Victoria Vanasco.

1.1 Introduction

When a firm approaches a bank to ask for a loan, the bank looks at the firm's observable characteristics to decide whether to approve the loan. It is very unlikely that these observables transmit all necessary information to evaluate how likely the firm is to default on the requested loan. One would expect that over time, if the loan is approved and subsequently monitored, the bank will learn something about the firm that was not reflected in the hard data provided with the initial loan application. In other words, through the process of establishing a relationship with the firm, the lender might obtain relevant but difficult-to-document "soft" information. By this we mean information that is qualitative in nature and consists mainly of ideas, opinions, rumors, feedback, or anecdotes which cannot be easily transmitted or verified by outside parties.

In line with this intuition, several studies have found evidence that borrower-lender relationships improve borrowers' access to credit. Research on relationship lending has shown that (i) there is something special about bank lending; and (ii) longer bank-firm relationships are correlated with cheaper access to credit. Slovin et al. (1993) examine the stock price of borrowing firms after the announcement of the failure of their main bank, Continental Illinois. They find that Continental borrowers incurred negative abnormal returns of 4.2% on average. If bank loans were indistinguishable from corporate bonds, borrowers could borrow directly from the market when their bank disappeared. Similarly, if banks were perfectly substitutable, the failure of one lender should have no impact on borrowers' stock prices. Slovin et al. conclude that Continental had private information about the borrowers unavailable to the rest of the market. Gibson (1995) reaches a similar conclusion by studying the effect of Japanese banks' health on borrowing firms. Petersen and Rajan (1994) and Berger and Udell (1995) independently show that a longer bank relationship (controlling for firm age) implies better access to credit in the form of lower interest rates or collateral requirements.

In this paper, we investigate what mechanisms result in a firm having better access to credit when it has established a relationship with a bank. Does establishing a relationship allow banks to receive soft information about borrowers? Does this learning occur only within a relationship – private learning – or are there spillovers to the market via public learning? To address these questions, we borrow the methodology developed by Farber and Gibbons (1996). These authors focus on learning and wage dynamics and show that time-invariant variables correlated with ability but unobserved by employers are increasingly correlated with wages as a worker's tenure increases. This evidence supports the idea that firms learn about worker quality over time. In this paper, we focus on interest rate dynamics and show that time-invariant variables correlated with firm fundamentals but unobserved by banks are increasingly correlated with interest rates over the course of a bank-firm relationship. Our results provide evidence that banks are able to privately learn about borrower fundamentals

in a way the market cannot.

We construct a panel of lender-borrower pairs ("relationships") observed repeatedly over time using the DealScan database on syndicated loans from Reuters LPC. DealScan provides detailed data for approximately 176,000 contracts comprising 248,000 syndicated loans made between 1981 and 2012. We match this extensive loan-level data with the financial characteristics of borrowing firms from the Compustat-CRSP Merged database. In our baseline loan pricing equations (similar to those developed in the banking literature), we show that even after controlling for observable borrower and loan characteristics, borrowers inside longer relationships pay cheaper loan spreads.

Why is there a discount for longer relationships? To test whether this is partially driven by bank learning about firm fundamentals, we construct a proxy for fundamentals which is not in the bank's information set. Our proxy is the differential response of the firms in our sample to a large negative aggregate shock: the recent financial crisis and the collapse of Lehman Brothers in September 2008. Specifically, we take the idiosyncratic component of firms' stock returns in the three months around the Lehman Brothers bankruptcy, and we orthogonalize it to all publicly-observable pricing variables at the beginning of each borrower-lender relationship in our sample, including the initial interest rate. For our identification strategy to work, the residual from this procedure must contain relevant pricing information about publicly-unobservable firm quality. We can be sure that banks are not learning directly about this proxy because the timing of its construction guarantees that it is never observed – the proxy is computed using future data. Moreover, the proxy is orthogonal to everything the bank used to price loans at the commencement of the relationship. Since we include the initial loan spread in the conditioning set, the orthogonalized proxy cannot be picking up the influence of omitted pricing variables. However, we find that the orthogonalize proxy is increasingly relevant for loan prices as a relationship progresses.

The orthogonalized proxy variable only contains information about firm fundamentals that were unobservable to each bank at the commencement of their relationship in our sample. We suggest that this information is correlated with private information that the bank acquires inside its lending relationship. To support this claim, we separately control for a second, public information proxy which only contains information about firm fundamentals that were unobservable to the market at the time each firm enters our sample. The private information proxy controls for each bank's initial information set, varying across relationships, while the public information proxy controls for the market's initial information set and only varies across firms. We find that even after controlling for market-wide learning about firm fundamentals over time, banks still price differentially on the private information proxy within a relationship. The relevant coefficient is 50 to 60% the magnitude of our initial estimate. This suggests that a significant portion of the value of bank lending is in private learning that occurs inside a relationship and is not shared by

all market participants.

The unique structure of our dataset allows us to control for time-varying firm characteristics. Since we observe multiple syndicates lending to the same firm during the same year, we are able to include firm-year fixed effects and control for any time-varying omitted variables which may have a non-stationary correlation with the private information proxy. In these fixed effect specifications we are holding all firm characteristics constant and comparing how two banks with two different length relationships price a loan. We find that the bank with a longer relationship puts greater weight on the private information proxy. Furthermore, the relationship length discount is negligible in this specification, hinting that the only reason why relationship lending matters is because of the transmission of soft information about firm's fundamentals.

In Section 2 we present a simple borrower-lender model and discuss the theoretical foundations of our empirical exercise. Section 3 is the main part of the paper. First, we discuss the nature of our dataset and the construction of our control and proxy variables. Second, we present our main empirical specifications and their results. Third, we present some robustness tests and discuss our results. Section 4 concludes.

1.2 A Simple Theoretical Framework

1.2.1 A Simple Model

Before describing our data and our empirical strategy, we present a simple model of firm borrowing to discuss the determinants of loan agreements, and the role of information in credit markets. We model a competitive banking system of risk-neutral banks with sufficient funds to finance all profitable projects. These banks have access to a risk-free rate R^F, which is exogenously given for an individual bank. Firms have insufficient funds to self-finance their heterogenous risky investment projects. Funds must be invested at the beginning of each period and payoffs are realized at the end of each period.

When a bank meets a firm that is demanding a loan, it determines the interest rate so as to be indifferent between lending to this firm or investing in the risk-free rate. Let I_0 be the information set of the bank when it first meets a firm in the market, at time 0, and let π be the probability of the firm defaulting on the requested loan. We assume π is the firm's private information and we denote the bank's beliefs about π at time 0 by $p_0 = E[\pi|I_0]$, i.e. the bank's expected probability of the firm defaulting on its loan, conditional on all available information at their first encounter. The bank determines the interest rate, R_L, given collateral, C_L, loan amount, L, and beliefs p_0 according to its own binding participation constraint:

$$(1 - p_0) LR_{L,0} + p_0 C_{L,0} = LR^F$$

Let $C_{L,0} = c_{L,0}L$, with $c_{L,0} \in \left(0, R^F\right)$ [2]. Let $r_{L,0}$ be the log excess return charged on a loan, i.e. $r_{L,0} = \log\left(R_{L,0} - R^F\right)$. We can re-write the pricing equation as follows:

$$r_{L,0} = \log\left(\frac{p_0}{1-p_0}\right) + \log\left(R^F - c_{L,0}\right)$$

This simple model predicts that the spread requested from a given loan increases with the expected default probability, p_0, and with the risk-free rate, and that it decreases with the percentage of the loan being collateralized. All of these results are standard and very intuitive.

We use this simple model to understand how the arrival of private signals about firm quality can affect the observed spreads on loans. If establishing a relationship with a firm allows the bank to observe private information about the firm's fundamentals, the bank should use this information to update its beliefs and recompute the required spreads.

Specifically, let $s^\tau = \{s_0, ..., s_\tau\}$ denote a time-series of i.i.d. private signals a bank receives during its relationship with a firm. The spread charged to the same firm for the same loan amount and same collateral after $\tau > 0$ periods will differ from the initial spread if s^τ is informative. Let $I_\tau = I_0 \bigcup \{s^\tau\}$. If signals are informative, $p_\tau = E[\pi|I_\tau] \neq E[\pi|I_0] = p_0$, and thus

$$r_{L,\tau} = \log\left(\frac{p_\tau}{1-p_\tau}\right) + \log\left(R^F - c_{L,\tau}\right)$$

Of course, this pricing equation might no longer be valid in the presence of asymmetric information in financial markets since the market is no longer perfectly competitive. We are indirectly assuming that all the surplus that arises from the bank-firm lending contract accrues to the firm. We could relax this assumption by adding an extra term that reflects how much of the reduction in interest rates goes to the borrower, and how much is exploited by the bank with private information.[3] What our model requires is that banks price to some extent on the arrival of private information, i.e., that the surplus arising from the relationship is shared. This is empirically the case.

In what follows we will decompose bank b's information set about firm f into three types of variables: $I_{fb,\tau} = \{x_{f,t_\tau}, z_{f,t_\tau}, s^\tau_{fb}\}$. The vector $x_{f,t}$ represents publicly-available characteristics of firm f at calendar time t which are observed by the bank but not by the econometrician (omitted variables). $z_{f,t}$ are public firm characteristics

[2]Note that if $c_{L,0} \geq R^F$, the loan would be made at the risk-free rate since even in default states the lender can get her outside option. Since we are interested in cases in which default does entails a loss for the lender, we focus on $c < R^F$.

[3]When this assumption is relaxed, the pricing equation is given by $r = r_{L,\tau} + f(\Delta, \gamma)$ where $f(\Delta, \gamma)$ is the share assigned to the bank, and depends on the lowest interest rate offerred by a competitor, $r_{L,\tau} + \Delta$, and on the firm's bargaining power, denoted by γ.

observed by both the bank and the econometrician (included variables). The set $s^{\tau}_{fb} = \{s_{fb,t_0}, s_{fb,t_1}, ..., s_{fb,t_\tau}\}$ represents the collection of private signals which only bank b observed during its relationship with firm f. The number of private signals is increasing in relationship length τ. For expositional purposes, suppose that firm characteristics $(x'_{f,t}, z'_{f,t})$ and other loan features $w_{l,fb,\tau}$ are time-invariant, so the "t" and "τ" subscripts may be suppressed.[4] We relax this assumption in the empirical section of the paper.

Consider a linearized version of the above pricing equation around the true default probability π[5]:

$$r_{l,fb,\tau} \approx \alpha_0 + \alpha_1 E[\pi | x_f, z_f, s^{\tau}_{fb}] + \gamma' w_{l,fb} \tag{1.1}$$

What if an econometrician could include the true default probability π in a panel regression along with observable characteristics $(z'_f, w_{fb,l})$? At relationship time 0, there would be a positive loading on π because of omitted variable bias: the bank's internal model includes variables x_f which are relevant for forecasting default probabilities and setting loan spreads. As a relationship progresses, the bank observes additional signals $s_{fb,t}$ which contain additional information about π not available in $\{x_f, z_f\}$. That is, the loading on π would increase over the course of the relationship due to private bank learning. This observation is at the heart of our empirical strategy.

1.2.2 Framework for the Empirical Strategy

Our aim is not to test this admittedly simple model but to use it as a motivation for our empirical specification. The core idea of our empirical strategy is taken from Farber and Gibbons (1996). These authors focus on learning and wage dynamics and show that time-invariant variables correlated with ability but unobserved by employers are increasingly correlated with wages as a worker's experience increases. In this paper, we instead focus on interest rate dynamics and show that time-invariant variables correlated by firms' fundamentals but unobserved by banks are increasingly correlated with interest rates as the bank-firm relationship increases. Our results provide evidence in favor of the idea that banks are able to learn over time about borrowers' fundamentals in a way the market cannot.

In our empirical model, we assume that the fth firm's default probability at time t follows an error-components structure which may depend on the macroeconomic environment m_t, industry-i-specific shocks v_i and idiosyncratic firm shocks $\xi_{f,t}$: $\tilde{\pi}_{f,t} := \eta_f + \tilde{\xi}_{f,t} = \eta_f + \alpha'_m m_t + v_i + \xi_{f,t}$. We allow for arbitrary forms of

[4]The "l" subscript on w counts if there are mutliple loans between the same bank-firm pair at the same point in time.

[5]For example, a first-order Taylor series expansion gives $r_{l,fb,\tau} = \frac{-\pi}{1-\pi} + \frac{1}{\pi(1-\pi)} p_{fb,\tau} + \log(R^F - c_{l,fb}) + o(p_{fb,\tau} - \pi)$.

cross-sectional and time-series correlation in the m_t and v_i components. These are nuisance parameters which may be removed by including time and industry fixed effects in our model, leaving two firm-specific components:

$$\pi_{f,t} := \eta_f + \xi_{f,t}$$

The parameter of interest to the bank as well as the econometrician is η_f, which we assume the bank does not know. We call this component a firm's latent *quality*. The following assumptions motivate our empirical strategy:

ASSUMPTION 1: There is a stationary distribution $F\Big(\eta_f,\xi_{f,t},x_{f,t},z_{f,t},b_f,s^{\tau}_{fb},m_t,v_i\Big)$ known by all bankers; i.e. bankers have symmetric information about the underlying distributions.

ASSUMPTION 2: Our dataset contains a time-invariant, background firm characteristic b_f which is correlated with η_f but has no direct effect on the probability of default: $E(\pi_{f,t}|\eta_f, b_f) = E(\pi_{f,t}|\eta_f)$.

ASSUMPTION 3: Non-interest contract features are conditionally uninformative about default probabilities: $E\left[\pi_{f,t}|x_{f,t}, z_{f,t}, s^{\tau}_{fb}, w_{l,fb,\tau}\right] = E\left[\pi_{f,t}|x_{f,t}, z_{f,t}, s^{\tau}_{fb}\right]$.

ASSUMPTION 4: Firm characteristics $(x'_{f,t}, z'_{f,t})$ are not informative about the idiosyncratic component of default probabilities: $E\left[\xi_{f,t}|x_{f,t}, z_{f,t}\right] = 0$.

ASSUMPTION 5: Default probabilities $\{\pi_{f,t} : t = 1, ..., T\}$ are cross-sectionally independent draws from a conditional distribution $G\left(\pi_{f,t}|\eta_f, x_{f,t}, z_{f,t}\right)$; i.e., shocks are conditionally i.i.d. across firms.

Unlike Farber and Gibbons, we assume that the information held by banks about firm quality is asymmetric. All banks know the distribution $F\Big(\eta_f,\xi_{f,t},x_{f,t},z_{f,t},b_f,s^{\tau}_{fb},m_t,v_i\Big)$, and the conditional distribution $G\left(\pi_{f,t}|\eta_f, x_{f,t}, z_{f,t}\right)$, all observe $\{x_{f,t}, z_{f,t}\}$ and whether a firm has defaulted or not, but they differ on their observed set of signals s^{τ}_{fb} as well as the number of signals (the length of the relationship) τ. The claim that we test in this paper is that access to these private signals allows the inside bank to price loans to firm f better than outside banks with a less-established relationship.

Imagine a panel dataset covering a cohort of firms entering the market for bank loans and taking out one-period loans from initially identical, perfectly competitive banks. The data reveal some firm and loan characteristics relevant for loan pricing ($z_{f,t}$ and $w_{l,fb,\tau}$, respectively) when the loan is applied for at the beginning of each period, but omits some firm characteristics $x_{f,t}$ relied on by the banks. Motivated by our linearized model (1.1), and given Assumptions 1-5, we could estimate the following population linear projection:

$$\begin{aligned} E^*[r_{l,fb,\tau}|z_{f,t}, w_{l,fb,\tau}] &= \alpha_t + \alpha_i + \alpha_1 E^*[E[\pi|x_{f,t}, z_{f,t}, s^\tau_{fb}]|z_{f,t}, w_{l,fb,\tau}] + \gamma' w_{l,fb,\tau} \\ &= \alpha_t + \alpha_i + \alpha_1 E^*[\pi|z_{f,t}, w_{l,fb,\tau}] + \gamma' w_{l,fb,\tau} \\ &= \alpha_t + \alpha_i + \beta^{z\prime} z_{f,t} + \beta^{w\prime} w_{l,fb,\tau} \end{aligned} \tag{1.2}$$

We use Assumption 3 to apply the Law of Iterated Linear Projections. The coefficient on w reflects both the substitutability between other loan characteristics and interest rate spreads (γ) and the correlation between w and omitted firm characteristics x and private signals s.[6] Similarly, the coefficient on z incorporates both direct and indirect pricing effects due to omitted variables.

Unobserved Firm Characteristics. b_f is a background firm characteristic in our dataset, but not observed by banks, that is correlated with latent firm quality η_f. We expect that b_f is unconditionally correlated with variables we omit in our pricing equation, $x_{f,t}$, that the bank uses in its forecast model $E[\pi_{f,t}|x_{f,t}, z_{f,t}s^\tau_{fb}]$. To remove this dependency, we use the residual from a regression of b on all observable firm characteristics and on the interest rate of the first loan in each relationship in our dataset. Conditioning on the latter ensures that b^*_{fb} is orthogonal to all the information held by each bank at the start of each relationship in our sample, including x_{f,t_0}. Specifically, let

$$b^*_{fb} = b_f - E^*\left[b_f|z_{f,t_0}, w_{l,fb,0}, r_{l,fb,0}\right] \tag{1.3}$$

This residual removes the influence of all information the bank may have used to price its first loan to a firm from the original background variable, b_f. Unlike the original background variable, b^*_{fb} may vary across banks for the same firm, so it carries an "fb" subscript.

Consider adding b^*_{fb} as a regressor to 1.2 with a slope which is allowed to vary over relationship time:

$$r_{l,fb,\tau} = \alpha_t + \alpha_i + \beta' z_{f,t} + \gamma' w_{l,fb,\tau} + \delta_\tau \cdot b^*_{fb} + \varepsilon_{l,fb,\tau} \tag{1.4}$$

We are interested in studying the evolution of the coefficient δ_τ. By the usual partitioned regression logic, if we define $b^\tau_{fb} = b^*_{fb} - E^*[b^*_{fb}|z_{f,t}, w_{l,fb,\tau}, t, i]$ as the residual from regressing b^*_{fb} on all other explanatory variables, then $\delta_\tau = Cov(b^\tau_{fb}, r_{l,fb,\tau})/Var(b^\tau_{fb})$ calculated cross-sectionally across firm-bank pairs at the same relationship time τ. By construction $\delta_0 = 0$. As banks receive additional signals s^τ_{fb}, private information becomes increasingly important in their internal forecast model $E[\pi_{f,t}|x_{f,t}, z_{f,t}s^\tau_{fb}]$. To the extent that b^*_{fb} is correlated with these private signals, the coefficient δ_τ should increase in magnitude with the number of signals and the length of the relationship τ.

[6] In our empirical specifications we find that the second factor dominates. For example, loans with more collateral pay higher interest rates, presumably because these firms differ on omitted characteristics.

In the next section we describe the construction of our dataset and how we test for private learning by constructing a time-invariant background variable b_f which is correlated with η_f but would have been impossible for banks to observe at the time the loans were made.

1.3 Empirical Analysis

In this section, we being by describing the dataset used for the empirical analysis. Next, we discuss our choice of a proxy variable for latent firm quality, b_f. Using this proxy, we proceed to test whether banks learn about customers as evidenced by an increasing loading on b_f within a specific lender-borrower relationship. We discuss and rule out several alternate explanations which might explain our finidngs, including public learning, time-varying omitted variables, and selection bias. Our results are consistent with the model described in the previous section. We find robust evidence that banks learn about unobserved firm characteristics while in a relationship.

1.3.1 Data

We construct a panel of lender-borrower pairs ("relationships") observed repeatedly over time. Specifically, we use the DealScan database on syndicated loans from Reuters LPC (April 2012 vintage). DealScan provides data for approximately 176,000 contracts comprising 248,000 syndicated loans made between 1981 and 2012, but the coverage between 1981 and 1987 is extremely limited; more than 99% of loans in the database start in 1988 or later. Syndicated loans are between a single borrower and a syndicate of lenders. One lender acts as the lead arranger and negotiates contract terms for the entire group. Most of the lenders are large commercial banks, but many syndicates include non-bank financial companies. After the contract is agreed to, a lender referred to as the agent monitors the performance of the loan. The lead arranger and agent can be different members of the syndicate. Each contract or "package" can include multiple loans or "facilities" made at the same time. A typical example is a borrower receiving both a term loan and a revolving line of credit.

Many of the rows in the DealScan tables contain missing values. The only filter we impose when tracking relationships over time is that lender and borrower IDs and deal dates are available, reducing our sample by approximately 3,000 facilities. For a given lender-borrower pair, we count every facility where that lender belongs to a syndicate lending to that borrower as an interaction in the relationship. There may be multiple observations at a particular moment in "relationship time" if a package contains multiple loan facilities. Since we care about the information set available to the lender at the time of the agreement, we order interactions by package date ("deal active date") rather than by each facility's specific start date. We restrict

our analysis to "lead arranger relationships," defined as bank-firm pairs in which the bank served as the lead arranger for at least one facility. Lenders playing an active role in arranging loan terms have greater incentives to acquire borrower information than passive members of the syndicate. In 56 percent of the relationships in our final sample, the lender served as the lead arranger in every interaction we observe with that borrower.

Our panel dataset requires information on loan prices and firm financial characteristics which the bank might use to set interest rates. Our measure of loan price is the all-included drawn spread over LIBOR, which is the price including fees that a firm would pay if it drew upon 100% of its line of credit (for revolving loans), and simply the spread over LIBOR including fees for term loans. Loans without an all-in spread are dropped. We obtain borrower financial data from Compustat using the link file created by Chava and Roberts (2008).[7] This reduces our sample by one half. Since our proxy variable is constructed from market data, we further require that the borrowers be publicly traded over the six-year period 2003-2008 and have stock return data available on CRSP (which we link using the CRSP-Compustat Merged database). Our data requirements restrict the sample to include only larger, more followed, and presumably more transparent firms. This should bias against finding any role for private bank learning. We drop all loans with a start date after 2003 to ensure the unobservability of our 2008-based proxy variable (see below), and we drop relationships in which the lender was never a lead arranger.

Our final dataset has 7,618 facilities and 5,740 relationships between 2,007 unique borrowing firms and 619 unique lenders. The deal active dates span the years 1987 to 2003. The average relationship lasts 3.5 interactions (approximately five years), and 10% of relationships last 7 or more interactions (approximately twelve or more years). Other summary statistics about the final sample of loans and relationships are provided in Table 1.1.

Observable Firm Characteristics

Our model requires that we condition on a subset of financial characteristics used by the bank in setting loan prices, $z_{f,t}$. Ideally these variables would be inclusive, so we do not have to worry about correlation between omitted variables $x_{f,t}$ and our proxy variable b_f (see the discussion below). While we could presumably condition on a laundry list of income statement ratios, we focus on a small subset of variables suggested in the literature on predicting corporate bankruptcies and defaults.

The oldest measure in this literature is Altman's Z score. Altman (1968) investigated the determinants of corporate bankruptcy for a sample of 33 manufacturing firms which filed for bankruptcy between 1946-1965 and 33 firms still in existence in

[7] We use the version of the link published on August 27, 2010, and made available on Wharton Research Data Services.

1966 based on random stratified matching by industry and size. He uses discriminant analysis to estimate the following index:

$$Z = (1.2 \cdot WC + 1.4 \cdot RE + 3.3 \cdot EBIT + 0.6 \cdot MVE + .999 * S)/AT$$

where WC is working capital, RE is retained earnings, EBIT is earnings before interest and taxes, MVE is market value of equity, S is sales, and AT is total assets.[8] Altman concludes that "firms having a Z score of greater than 2.99 clearly fall into the 'non-bankrupt' sector, while those firms having a Z below 1.81 are all bankrupt" (p. 606). So lower values of Z indicate an increased likelihood of bankruptcy. We winsorize the top and bottom 0.5% of Z-score observations using the sample of all DealScan firms for which we have data over the years 1985-2012.

Our second measure comes from the observation in Merton (1973) that the Black and Scholes (1973) options pricing model may also be used to calculate the market value of assets in place, by viewing the observed equity price as a call option on the unobserved market value of the entire firm. Once the market value of assets in place V_A has been estimated, a firm's probability of default T periods into the future is the probability that the value of its assets will drift below the "strike" price–the book value of liabilities. Since the Merton model assumes that V_A follows a geometric Brownian motion with deterministic drift μ and volatility σ_A, this probability is given by

$$P(V_{A,t+T} \leq L_t | V_{A,t}) = \Phi\left(-\frac{log(V_{A,t}/L_t) + (\mu + \frac{1}{2}\sigma_A^2)T}{\sigma_A\sqrt{T}}\right)$$

To calculate this exact probability, one must solve the Black-Scholes equations for V_A and σ_A. Rather than using a numerical solver, we use the "naive" alternative proposed by Bharath and Shumway (2004, 2008). This naive probability of default uses simple rules of thumb for variables in the formula above: L_t is the book value of debt in current liabilities plus one-half the book value long-term debt; V_A is the sum of market value of equity plus book value of liabilities; equity volatility σ_E is the annualized standard deviation of the previous year's daily stock returns; debt volatility $\sigma_L = .05 + .25 \cdot \sigma_E$; and total firm volatility is the weighted sum of σ_E and σ_L. We solve for the naive probability of default for firm f at time t, $NPD_{f,t}$ for a one-year time horizon. In all tables and regressions, we truncate the probability of default to take values in the range $[0.001, 0.999]$.

Our observable firm characteristics which are relevant for loan pricing are thus two measures for predicting corporate bankruptcy or default on debt obligations: $z_{f,t} = (Z_{f,t,}NPD_{f,t})'$.

[8]There is an error in the placement of a decimal point in the original 1968 paper. The correct formula is given in subsequent papers–e.g., Altman (1984).

Construction of the Private Information Proxy

A good background variable b_f cannot be in the bank's information set at any time and it must be correlated with the firm's unobservable latent quality, η_f. Our candidate background variable is the differential response of the firms in our sample to a large negative aggregate shock: the onset of the financial crisis and the collapse of Lehman Brothers in September 2008. Specifically, we consider the idiosyncratic component of firms' stock returns in the three months around the Lehman Brothers bankruptcy. By using equity market data from five years after the last loan in our sample was made, we guarantee that the proxy cannot have been observed by banks in real time. Lehman's bankruptcy filing was a "shock" in the sense that it was not foreseen by market participants and triggered a re-evaluation of expected returns on investments across the entire economy. When Bear Stearns failed six months earlier, the Fed and the Treasury avoided the bankruptcy process and arranged its purchase by JP Morgan Chase precisely to ameliorate turmoil in financial markets.

We require that idiosyncratic stock returns around the Lehman filing were partially driven by firms' latent ability. Suppose that during booms it is hard to differentiate good firms from bad firms, while during busts lemons are easier to identify. Those firms that perform relatively better during crises are spotted as high-quality firms, and investors should incorporate this information into the stock price. Moreover, the returns to identifying lemons might be greater in crisis states of the world; in booms all firms do well, while in busts only good firms do well. If signals about firm quality became more informative after Lehman, or if investors' incentives to acquire costly information increased, then the main news content in the months after this shock should be a reassessment of firm quality. Of course, a component of firms' stock returns during this period undoubtedly reflect subprime-crisis-specific exposure. To the extent that subprime exposure is industry-specific, we can remove this influence with industry fixed effects. Our identifying assumption is that at least part of firms' idiosyncratic returns are due to underlying firm characteristics that were revealed after Lehman, and not to subprime-crisis-specific risk exposure. We do not interpret loadings on the proxy as changes in the perceived probability of a Lehman-style crisis occurring, as we find it implausible that this risk was priced in loans made a decade or more in advance.

We construct b_f as follows. We compute the cumulative abnormal return of each firm in a [-21, +42] day window centered around the collapse of Lehman:[9]

$$b_f := \sum_{s=-21}^{+42} \left(R_{f,s} - R^F \right) - \hat{\beta}_f' \left(R_{factor,s} \right)$$

where $R_{f,s}$ and $R_{factor,s}$ denote the daily returns on a firm's stock and the four Fama and French (1993) - Carhart (1997) factors at time s, R^F denotes the risk-free rate,

[9]Starting on August 14 and ending on November 12.

and $s = 0$ is September 15, 2008. The factor betas are estimated from time-series regressions of daily excess stock returns over 2003-2007:

$$R_{f,t} - R^F = \alpha_f + \beta_f' \left(R_{factor,t}\right) + \varepsilon_{f,t}$$

With each firm's CAR in hand, the final **private information proxy** is given by (1.3). We define relationship time 0 as the time of the first loan between a firm-bank pair in our sample. It is likely that the time of first observation is not the first interaction between a bank and firm for many loans. Nevertheless, by orthogonalizing at the first non-censored observation, we can remove the influence both of omitted variables and of any private learning that may have occurred within the censored relationship observations. To the extent that learning is diminishing over time, the inclusion of mature relationships will bias our estimates toward zero.

The orthogonalization guarantees that b^*_{fb} is uncorrelated with relevant omitted firm characteristics at the start of each relationship, x_{f,t_0}. However, a failure of Assumption 4 would pose an identification problem if the idiosyncratic component of default probability $\xi_{f,t}$ and omitted variables $x_{f,t}$ jointly exhibit within-firm autocorrelation. That is, since b^*_{fb} is from the future, the proxy could simply be picking up future innovations in a firm's default probability which are correlated with subsequent movements in publicly available variables. The unique structure of our panel dataset, in which we observe the same borrower in different relationships at the same period in calendar time, will allow us to resolve this problem by applying firm-year fixed effects

The coefficients from the orthogonalization regression are presented in the first column of Table 1.2. Note in particular that the all-in-spread at time zero is negatively correlated with the Lehman proxy, even after controlling for Z score, naive probability of default, other loan characteristics, and industry fixed effects. A firm paying an additional 100 basis points on its first loan in our dataset is expected to experience an additional 2.7 percentage point negative CAR in the three-month window around Lehman. This indicates that initial loan prices contain omitted information which is correlated in the correct direction with the proxy variable. The private information proxy b^*_{fb} is simply the residual from this regression.[10]

1.3.2 Testing for Bank Learning

We begin the main part our anaysis by estimating a standard pricing equation, to be sure that our data replicates results already highlighted in the literature. We regress the all-in drawn spread of each loan on firm and loan characteristics, and on relationship time:

$$r_{l,fb,\tau} = \alpha_t + \alpha_i + \beta' z_{f,t} + \gamma' w_{l,fb,\tau} + \varphi \cdot \tau + u_{l,fb,\tau} \tag{1.5}$$

[10] If the initial package contained more than one facility, we include in the regression all loans in that package. The private information proxy is then the average of the residuals: $b^*_{fb} = 1/L \sum_{l=1}^{L} b^*_{l,fb}$.

where each observation is given by a loan l between firm f and bank b at relationship time τ. We control for year t and two-digit SIC industry i with fixed effects. Results are presented in the second column of Table 1.2. Larger predicted probabilities of default (lower Z score and higher NPD) are associated with higher spreads, while longer relationships are associated with a discount in the spread equal to 3.7 basis points per interaction. Secured loans have on average higher spreads, a seemingly counterintuitive result. This and other loan characteristics are likely reflecting some unobservable characteristic that the bank is pricing. If secured loans are of worse quality on unobservables, then they should pay higher spreads. Finally, we find that longer-term and revolver loans are associated with higher interest rates (although the coefficient on loan maturity is not statistically significant).

The main result from this regression is that having an established relationship with a bank lowers the cost of credit for a firm even after controlling for relevant pricing characteristics. The effect is independent of a borrower's quality, as measured by Z score and NPD. We proceed to test whether this relationship discount is due to unobserved learning or something else.

In our baseline learning specification, we add the private information proxy b^*_{fb} to the previous regression. By construction the proxy variable can have no effect on loan prices at relationship time zero. The test is whether the loading varies over relationship time and whether "better" firms receive a discount. The coefficient of interest is δ_τ in the following specification:

$$r_{l,fb,\tau} = \alpha_t + \alpha_i + \beta' z_{f,t} + \gamma' w_{l,fb,\tau} + \delta_0 \cdot b^*_{fb} + \delta_\tau \cdot \left(b^*_{fb} \times \tau\right) + \varphi \cdot \tau + u_{l,fb,\tau} \quad (1.6)$$

Estimates are presented in the third column of Table 1.2. First note that the inclusion of our proxy variable does not affect any of the results obtained in the baseline case. Second, the coefficient on the proxy variable interacted with relationship time has a highly significant effect on the pricing of a firm's loans. Consider a one standard devation increase in the proxy, an increase in the CAR of 0.37 log units (i.e., 37 percentage points). Holding other firm and loan features constant, this firm would benefit from a reduction in its interest rate on bank loans of (-5.337)·(0.37) = -1.96 basis points per renewal. On an average sized loan ($358 million), this would result in annual savings of $70 thousand per year. Since the average maturity of a loan in our sample is just over four years, the total savings from renewing its loan with an existing lender instead of switching lenders is $280 thousand for the first renewal. The savings increases with relationship length: on the fifth renewal it would be $1.4 million.[11] Put another way, a one S.D. increase in the proxy has the same benefit per renewal on loan prices as a 1.4 percentage point decrease in the Merton-Bharath-Shumway naive probability of default.

We conclude from this regression that the proxy variable is correlated with information that banks use to price loans. Furthermore, the banks did not have this

[11] This savings is about half the magnitude of the baseline relationship effect, a discount of 3.7 basis points per renewal.

information at the time of the first loan in our sample. In the next section we test whether the effect is unique to banks that have a relationship with a firm. In other words, is learning public or private?

1.3.3 Public vs. Private Learning

The previous regression has shown that banks act "as if" (to quote Milton Friedman) they price loans on what we have referred to as a private information proxy. This proxy derives from stock market returns in the second half of 2008, while the most recent loan in our sample is from August 2003, so banks cannot have actually priced on this proxy. This suggests that information correlated with both the proxy and latent firm quality is revealed to market participants as relationship time increases. However, we have not ruled out the alternate explanation that learning is public. That is, it is possible that banks learn about firm quality over time, but this information is non-excludable and the benefits diffuse across all lenders. To distinguish between private and public learning, we need access to a second proxy which only contains information about firm fundamentals that were unobservable to the market at the time of a firm's first syndicated loan in our sample, t_{00}. By being orthogonalized to information available to the market at the time the firm enters our sample, this proxy should reflect any pricing based on public information. We construct such a **public information proxy** as follows:

$$b_f^* = b_f - E^*\left[b_f | z_{f,t_{00}}, w_{l,fb,00}, r_{l,fb,00}\right] \tag{1.7}$$

The public information proxy only varies across firms, not across relationships. To the extent that learning about firm quality is public, the loading of interest rates on the public information proxy should increase with the time that the firm has been present in the market. If all bank learning about firm quality is public, then the loading on the private information proxy within a specific bank-firm relationship should drop out once we control for market-wide learning.

To implement this test we estimate the following regression equation:

$$\begin{aligned} r_{l,fb,\tau} &= \alpha_t + \alpha_i + \beta' z_{f,t} + \gamma' w_{l,fb,\tau} + \delta_0 b_{fb}^* + \delta_\tau \cdot \left(b_{fb}^* \times \tau\right) + \delta_{00} b_f^* \\ &+ \delta_t \cdot \left(b_f^* \times (t - t_{00})\right) + \varphi_\tau \cdot \tau + \varphi_t \cdot (t - t_{00}) + u_{l,fb,\tau} \end{aligned} \tag{1.8}$$

Results are presented in the first column of Table 1.3. The estimated value of δ_τ is -2.7 and of δ_t is -2.8. Both coefficients are about half the magnitude of our baseline estimate of -5.3 from Table 1.2, column 3, and both are significant at smaller than the 5% level. These results suggest that banks outside a relationship do in fact learn about the firm's quality over time. This may be due to the evoluion of observable fundamentals that we omit from our pricing equation ($x_{f,t}$). It could also indicate that outside banks are able to partially infer the inside bank's private information from publicly-observable signals such as the terms of loan renewals. However, even

after controlling for the possible presence of market-wide learning, we continue to find a large and statistically significant loading on the private information proxy. Banks inside a relationship are able to price on firm quality differentially from banks outside a relationship. This is strong evidence in favor of our argument that information about firm quality is privately transmitted inside the bank-firm relationship.

1.3.4 Alternate Explanations

Forecast Window Effect

One potential confounding factor is that our private information proxy is taken from future financial market data. It might be the case that all market participants are forecasting some factor correlated with b^*_{fb}, such as future earnings, and that these forecasts mechanically become more accurate as $t \to 2008$ simply because the forecast window is shrinking. To be confounding, such an effect would have to manifest as an interaction between the private proxy and calendar time. If there were something special merely about time until 2008, it would be picked up by the calendar year fixed effects. Furthermore, we have already controlled for market-wide pricing on the public component of our background variable in Table 1.3, column 1. An important component of loan pricing specifically appears to occur inside a relationship, which is evidence of private bank learning.

As a robustness test, we re-run regression (1.8) with the private information proxy interacted with indicator variables for each year. This specification should remove any mechanical correlation between the private information proxy and loan rates which depends on calendar time but is independent of relationship time, such as a forecast window effect. The estimates from this specification are presented in the second column of Table 1.3. The results are very similar to our tests for public learning and do not alter our finding that the private proxy interacted with relationship time is an important factor in the bank's pricing decisions.

Omitted Firm Variables

So far we have assumed that the banks can only learn about the permanent component of default probability η_f. This comes from Assumption 4, that firm characteristics are uninformative about the idiosyncratic component of default probabilities $\xi_{f,t}$. A plausible alternative assumption is that firm characteristics and idiosyncratic shocks $(\xi_{f,t}, z'_{f,t}, x'_{f,t})$ exhibit contemporaneous correlation, for example due to a common driving process or a triangular VAR structure. It can be shown that if $\xi_{f,t}$ exhibits serial correlation, then the magnitude of $Cov(b^*_{fb}, x_{f,t})$ is increasing in t. Intuitively, the non-orthogonalized background variable contains information about both the total default probability and omitted firm characteristics in 2008. The orthogonalization procedure removes the influence of omitted variables at relationship time 0 but leaves information about total default probability. If subsequent values of

x contain information about subsequent innovations in the default probability, this will show up as a correlation with the orthogonalized private information proxy. As the innovations accumulate, the correlation will increase in magnitude. This will exhibit as omitted variable bias in our regressions – we would mistake banks pricing on publicly-observable variables for private learning.

Our data includes multiple banks lending to the same firm during the same calendar year, so it is possible to control for firm-by-time omitted variables $x_{f,t}$ using firm-year fixed effects. Within a firm-year, two banks should price loans differently only if they have access to different private information, s^{τ}_{fb}. This test is very stringent: the fixed effects alone absorb over 96% of the variation in the all-in spread.[12] The remaining variation comes from banks in different syndicates lending to the same firm f in the same calendar year t but with different length relationships τ. The coefficients on relationship time and its interaction with the private information proxy are identified from this remaining variation.

Firm-year fixed effect results are presented in the third column of Table 1.3. The firm-year fixed effects absorb the public information proxy and its interaction with market time, so coefficients on those variables are not shown. The fixed effects absorb most but not all of the variation in the annual firm controls – these variables do not drop out due to heterogeneity in fiscal year-end dates. In particular, the market-based NPD remains significant and similar in magnitude to previous equations.

The coefficient of interest to us is the interaction between the orthogonalized proxy and relationship time. Even in this very demanding specification, the coefficient remains statistically different from zero. The magnitude is about a quarter as big as our baseline estimate: a one S.D. increase in the CAR is now associated with a half basis point discount per renewal. We note with some suprise that relationship length is by itself economically small, not significant, and the wrong sign. This suggests that after controlling for all possible firm characteristics, the only remaining channel through which relationships matter is the transmission of private information.

Our theory once again passes the test: the private information proxy is not merely capturing some publicly-observable, omitted firm characteristic that varies over time. It suggests that within a relationship, banks receive private information that allows them to better estimate firm quality, and that this information is used when pricing a firm's loans.

Other Possible Explanations

In this subsection we discuss other possible explanations for our results.

Functional form misspecification. Suppose the true pricing equation is a non-linear function of firm characteristics z, and that the proxy variable is correlated

[12] Also, firms which take out loans from different syndicates in the same year may differ systematically from firms which take out loans with only one syndicate in the same year.

with this non-linear function. Controlling for z in a linear fashion is misspecified and does not remove the relevant correlation. However, any spurious relationship between b^*_{fb} and $r_{l,f,\tau}$ should be constant over time. This does not explain our result that the loading on the proxy increases with relationship time.

Selection bias. Suppose that banks screen on omitted but publicly-observable firm characteristics x, so that only the best firms have long-term relationships. In the extreme case, imagine that there are two firms, G and B. Firm G stays in a long-term relationship with its bank and pays a low interest rate because it is high quality, while firm B switches banks every period and pays a high interest rate because it is low quality. This would create a negative correlation between relationship length and interest rate spreads in our data. However, we control for relationship length and find that the interaction between relationship length and the proxy variable also matters.

Reverse causality. It might be the case that firms with longer relationships had easier access to funds during the credit crunch surrounding Lehman, enabling them to better weather the shock. If firms in longer relationships receive lower interest rates for reasons unrelated to bank learning, we could find a spurious correlation between interest rates and the Lehman CAR which is increasing in relationship length. To address this point, we compute the correlation between the Lehman CAR and the length of a firm's longest active banking relationship in December 2003, the last year of our sample. We label a relationship as "active" if the most recent loan in the relationship either matured after November 2001 or was still in place. The results are presented in Table 1.4. We find a weak correlation between a firm's longest relationship and its CAR to Lehman, but the effect is only significant (at the marginal 10% level) when we include both active and inactive relationships. We find no statistically significant correlation between a firm's time in the market or its longest active relationship and its response to Lehman. Moreover, the R-squared from all three specifications is essentially zero, indicating that any possible role for reverse causality is extremely small.

1.4 Conclusions

We began this paper by posing the question, "Do banks learn?" Our answer is a resounding yes. We first verified that borrowers inside longer relationships pay cheaper loan spreads, as previously shown in the literature of relationship lending. We then tested whether this reduction in spreads could be partially driven by banks learning about firm fundamentals using the methodology developed in Farber and Gibbons (1996). We constructed a proxy for firm fundamentals which is orthogonal to the bank's information set, based on the differential response of the firms to the collapse of Lehman Brothers in September 2008. We argue that this contains relevant information about firm's tail risk, which is precisely what lenders care about when

pricing loans in this market. We showed that our proxy is increasingly relevant for loan prices as a relationship progresses. Even after controlling for market-wide learning about firm fundamentals over time, banks still price differentially on the private information proxy within a relationship.

In future research, we plan to further investigate what it is that banks are learning about. Possible candidates include: (i) firm-specific characteristic, such as the real value of assets in place, or the effectiveness of the firm's corporate governance structure; (ii) the top management's character and ability; or (iii) the membership and the activeness of the firm's board. We will exploit variation in CEOs and board membership across firms to disentangle these possible explanations.

Table 1.1: Summary Statistics

Relationships where lender is sometimes the lead arranger (NT_n = 13,954). Panel includes 2,007 unique borrowers and 619 unique lenders.

	N	Mean	SD	10th Percentile	90th Percentile
Panel A: Relationship Characteristics [1]	**5740**				
Length (# of interactions)		3.50	2.88	1	7
Calendar length (months)		57.6	61.7	1	143
Fraction always lead arranger		0.542			
Fraction sometimes agent		0.993			
Fraction always agent		0.650			
Panel B: Loan Characteristics [1]	**7618**				
All-in spread (bps)		138	116	25	300
Loan Size ($m)		358.4	727.9	10.0	864.0
Maturity (months)		36.4	24.8	12	60
Fraction revolver		0.630			
Fraction collateralized		0.347			
Fraction not collateralized		0.231			
# times appears in relationship panel		1.8	1.2	1	3
Panel C: Borrowing Firm Characteristics [1]	**5509**				
Total assets ($b)		10.803	52.077	0.075	16.851
Average Q [2]		1.376	1.222	0.476	2.572
ROA (%)		3.260	11.598	-2.735	11.322
Z score		2.466	1.713	0.659	4.544
Naïve Probability of Default (% / 100)		0.056	0.179	0.001	0.114
Three-month CAR around Lehman (% / 100)		-0.082	0.367	-0.568	0.322
# facilities per borrower-date [3]		2.6	2.4	1	5

Notes.

[1] In Panel A, each observation is a lender-borrower pair; in Panel B, a facility; in Panel C, a firm-date.

[2] Q = (E + P + D) / A, where E is market value of common equity, P is liquidating value of preferred stock, D is book value of long-term debt plus current liabilities net of (current assets less inventories), and A is book value of total assets.

[3] Count includes multiple facilities per package and multiple packages taken out in same fiscal period.

Table 1.2: Do Banks Learn?

OLS Panel regression of bank-firm "relationships" over time.

Dependent variable:	*Lehman Proxy*	*Interest Rate spread over LIBOR (in bps)*	
	(1)	(2)	(3)
All-in Spread at Rel. Time 0	-0.000268***		
	(0.000)		
Borrower's Z score	0.0228***	-9.674***	-9.611***
	(0.004)	(0.918)	(0.924)
Naïve Probability of Default	-0.0077	143.4***	142.9***
	(0.027)	(9.741)	(9.612)
Relationship Time		-3.731***	-3.743***
		(0.577)	(0.571)
Private Info Proxy (see note)			3.269
			(4.044)
Private Info. Proxy * Relationship Time			-5.337***
			(1.361)
Total Assets ($b)	-0.000103*	-0.151***	-0.152***
	(0.000)	(0.026)	(0.026)
1{loan is secured}	-0.0263*	109.0***	109.0***
	(0.013)	(3.585)	(3.610)
1{loan is not secured}	0.00952	-1.54	-1.596
	(0.010)	(2.068)	(2.074)
Loan Maturity (months)	-0.000598***	6.62E-02	6.63E-02
	(0.000)	(0.046)	(0.046)
1{revolver loan}	0.0157*	6.697***	6.591***
	-0.00931	-2.073	-2.081
Year FX	YES	YES	YES
Industry FX	YES	YES	YES
Observations	7,390	13,954	13,954
R-squared	0.238	0.489	0.49

Standard errors clustered by lender in parentheses.
*** p<0.01, ** p<0.05, * p<0.1

Notes.

The "Lehman proxy" is the 3-month cumulative abnormal return from a Fama-French-Carhart four-factor centered around the Lehman bankruptcy of 9/15/2008.

Column 1 reports a cross-sectional regression of the proxy on all dependent and independent variables as of the first interation between each borrower-lender pair in our sample (relationship time 0). This may include multiple facilities per relationship.

Colums 2 and 3 uses the residuals from Column 1 as the "orthogonalized" Private Info Proxy. This proxy is re-calculated whenever a borrower changes lenders.

Table 1.3: Is Learning Private or Public?

OLS Panel regression of bank-firm "relationships" over time.

Dependent variable:	*Interest Rate spread over LIBOR (in bps)*		
	(1)	(2)	(3)
Borrower's Z score	-9.625***	-9.668***	-6.306
	(0.859)	(0.865)	(5.885)
Naïve Probability of Default	138.8***	137.5***	164.8***
	(9.489)	(9.402)	(57.290)
Relationship Time	-1.953***	-1.850***	0.0424
	(0.502)	(0.495)	(0.131)
Years in Market	-2.391***	-2.481***	*absorbed*
	(0.275)	(0.286)	*by FX*
Private Info Proxy	66.49**	*absorbed*	79.46***
	(26.610)	*by FX*	(19.130)
Public Info Proxy	-54.63**	*absorbed*	*absorbed*
	(26.660)	*by FX*	*by FX*
Private Info Proxy *	-2.703*	-3.153**	-1.440**
Relationship Time	(1.430)	(1.489)	(0.568)
Public Info Proxy *	-2.794***	-2.153***	*absorbed*
Years in Market	(0.705)	(0.789)	*by FX*
Total Assets ($b)	-0.157***	-0.152***	-0.165
	(0.026)	(0.026)	(0.169)
1{loan is secured}	107.3***	106.9***	16.14***
	(3.482)	(3.427)	(4.857)
1{loan is not secured}	-0.276	-0.4	-12.29***
	(1.984)	(1.944)	(2.472)
Loan Maturity (months)	0.0608	0.0597	0.108***
	(0.046)	(0.046)	(0.028)
1{revolver loan}	6.347***	6.371***	-5.718***
	(2.047)	(2.033)	(1.320)
Year FX	YES	YES	YES
Industry FX	YES	YES	YES
Proxy*Year FX		YES	
Borrower*Year FX			YES
Observations	13,954	13,954	13,954
R-squared	0.497	0.502	0.958

Standard errors clustered by lender in parentheses.
*** p<0.01, ** p<0.05, * p<0.1

Note.

The proxy variables are constructed from a 3-month cumulative abnormal return in a Fama-French-Carhart four-factor model centered around the Lehman bankruptcy of 9/15/2008.

The Private Info Proxy is re-orthogonalized to all dependent and independent variables at the beginning of each relationship (Table 2 col. 1). The Public Info Proxy is orthogonalized only once: the first time the firm enters the market.

Table 1.4: Do Long Relationships Predict *b* ?

OLS cross-sectional regression of borrowing firms in 2003.

Dependent variable:	*Lehman Proxy*		
	(1)	(2)	(3)
Longest Active Relationship as of 12/2003	0.00432		
	(0.003)		
Longest Relationship in Sample as of 12/2003		0.00583*	
		(0.003)	
Years in Market as of 12/2003			0.00134
			(0.002)
Constant	-0.100***	-0.105***	-0.0990***
	(0.011)	(0.012)	(0.014)
Observations	2002	2002	2002
R-squared	0.001	0.001	0.000

Robust standard errors in parentheses

*** p<0.01, ** p<0.05, * p<0.1

Chapter 2

Distance, Asymmetric Information, and Mortgage Securitization

2.1 Introduction

Why do originating lenders sell some mortgages for securitization and not others? This question goes to the heart of what role the market for mortgage-backed securities played in the financial meltdown of 2007-08. The mortgage lending and securitization process involves many successive layers of asymmetric information (**?**, Dai et al. 2013). First, the homeowner knows her own intentions, while the originating loan officer must forecast her default and prepayment probabilities from information in her credit report and loan application. Second, there is an agency problem between the loan officer who carries out lending policy and his supervisors who set the institution's policy and must monitor his performance (Stein 2002). Third, the originating institution knows more about the borrower and her economic situation than any potential purchaser can. The lender's informational advantage should be particular strong for local loans. A local lender has specialized knowledge about economic conditions in its home geography and benefits from informational spillovers with existing customers (Winton 1999, Garmaise and Natividad 2013). A local lender is more likely to extend a "character loan" and rely on soft information when making its credit decision (Liberti and Mian 2009). Moreover, the borrower is more likely to have a prior relationship with a local lender, due to search costs and the benefits of relationship lending (Petersen and Rajan 1994, Berger and Udell 1995, Degryse and Ongena 2005).

This paper argues that the informational advantage of the originating institution, relative to a potential purchaser of the mortgage, strongly affects the selection of mortgages that are offered for sale. The stronger the asymmetry, the less likely it is that the mortgage can be sold, in the spirit of a classic lemon problem. However, it is not *ex ante* clear that the lemon problem is so dominant. More private information also implies that the lender is likely to have picked or attracted borrowers of higher average quality. That is, lower information acquisition costs for the lender ease the purchaser's monitoring problem and could increase the likelihood of trade taking place. I illustrate this tension in a simple game-theoretic model, which shows that either the "lemon" effect or the "quality" effect could dominate. Turning to the data, I show that the lemon effect dominates – lender private information hinders trade on secondary markets. I consider and rule out a number of plausible alternative explanations, including different lender business models, purchaser diversification needs, and foreclosure costs.

I consider a setting where lenders can acquire private information about borrowers that they cannot credibly communicate to secondary-market participants. This creates a lemon problem in secondary markets: any loan being offered for sale is probably of below-average quality. This intuition fits with prior evidence that secondary mortgage market sellers use their private information to the disadvantage of buyers (Downing et al. 2009, Agarwal et al. 2012). Frictions that increase the cost of lender learning can alleviate this problem. The geographic distance between

borrowers and lenders is such a mechanism. Berger et al. (2005) document that the informational intensity of small business loans varies inversely with lender-borrower distance: more distant loans involve less frequent and less personal forms of communication, so there is less opportunity for lenders to acquire so-called "soft" information. Greater lender-borrower distance dampens the lender's informational advantage, so it should be easier to sell mortgages from faraway borrowers.

Of course, borrowers could anticipate that lenders exert less effort originating faraway mortgages, particularly if they plan on selling the loans. This might lead to an offsetting "quality effect": as lender-borrower distance increases, the quality of the average mortgage unobservably declines. Previous research has shown that the "originate-to-distribute" (OTD) model worsens agency problems and reduces bank incentives to exert effort in the origination process. Keys et al. (2010) exploit a discontinuity in the ease of securitizing mortgages when the borrower has a FICO score just above 620 and show that these loans default 10-25% more often. Purnanandam (2011) finds that banks engaging in more OTD lending during the boom experienced higher default rates on their real estate portfolio after secondary markets dried up, evidence that banks planning to sell their loans did not expend resources in screening their borrowers. More tellingly, La Cava (2013) finds that the retained mortgage loans of banks which do more distant lending non-perform at higher rates. Since lender originating effort is unverifiable, secondary-market purchasers must weigh whether the decline in average borrower quality is big enough to offset the decline in probability that a more distant loan is a lemon. The mortgage sale rate might either rise or fall with lender-borrower distance, depending on which effect dominates.

I test the importance of these two channels using Home Mortgage Disclosure Act data for MSAs in the continental United States over 1990-2000, a "normal" period that predates the housing boom and bust. This dataset provides near-comprehensive coverage of lender identities, borrower location at the county level, loan amount, approve-or-deny decisions, and hold-or-sell outcomes for the residential mortgage market. The dataset does not provide any price information prior to 2004, so my results focus on quantity-based measures.

This paper is the first to show a systematic link between distance and mortgage sale rates: as distance between the borrower and the lender's headquarters location increases, the mortgage sale rate increases. This is suggestive evidence that bankers acquire and rely on soft information even when originating a highly-standardized product such as a mortgage loan. This finding complements previous theoretical and empirical evidence that lenders face higher information acquisition costs for more distant borrowers, leading to worse *ex post* outcomes but also lower information asymmetries vis-à-vis secondary market purchasers (Frankel and Jin 2011, La Cava 2013). My second contribution is to interpret this finding using a strategic model of lender-purchaser interaction. The model indicates that asymmetric information in the first two links of the mortgage securitization chain, from borrowers to originators and from originators to purchasers, push in opposite directions. The positive coef-

ficient on distance indicates that market participants themselves act as though the second asymmetry is more severe – i.e., that the lemon effect dominates the quality effect.

While I interpret this finding in terms of asymmetric information – demand is increasing with distance, because purchasers believe distant loans are less likely to be lemons – there are other possible explanations. For example, large lenders have different business models than small lenders, so they might do more distant lending and sell a larger fraction of their overall portfolios. Another alternative explanation is that lenders face higher costs in managing the foreclosures of more distant properties, so the supply of loans being offered for sale is increasing with distance. To help pin down the link between distance and asymmetric information, I delve into the unique role played by two government-sponsored institutions in the secondary mortgage market. Fannie Mae was established in 1938 and Freddie Mac in 1970 to promote access to low-interest rate, long-term home mortgage loans by allowing banks to shed some of the duration and prepayment risk associated with holding these loans (Green and Wachter 2005). There are strict guidelines to which mortgages must adhere to be eligible for delivery to Fannie and Freddie, foremost of which is a nationally-set dollar cap. The Conforming Loan Limit segments the mortgage market in two: the highly liquid conforming-size market below the cap, and the much less liquid jumbo-size market above the cap. An originating lender faces much greater incentives to acquire information about a jumbo-size mortgage than about a conforming-size mortgage, because the jumbo loan is less likely to be successfully sold.[1]

Given this discrete difference in seller incentives, I predict that the effect of distance on a mortgage's probability of being sold should be larger for a jumbo mortgage than for an otherwise-comparable conforming mortgage. I analyze the mortgage sale rate to private counterparties for loans within $10,000 on either side of the CLL and test for a discontinuity in the distance coefficient at the CLL. The private sale rate increases by 2 to 3 percentage points for every doubling of borrower-to-lender-HQ distance for loans taken out just above the CLL. There is no robust statistical relationship between distance and private sale rates for loans taken out just below the CLL. My controls include lender×year fixed effects, so the finding is not driven by different lender business models. Moreover, I present evidence that the discontinuity occurs exactly at the CLL and, although borrower sorting around the CLL is present, the discontinuity in the distance coefficient is not driven by borrower sorting. This is further evidence that the positive link between lender-borrower distance and the mortgage sale rate is due to diminishing lender-purchaser informational asymmetries and not some other mechanism.

[1]Loutskina and Strahan (2009) previously exploited this discontinuity to study the effects of banks' financing constraints on credit supply. They find that bank financial conditions affect the supply of loans just above the CLL but not just below the CLL, measured in terms of both loan volume and acceptance rates. This is evidence that easier loan securitization dampens the impact of local bank shocks on the credit decision.

Frictions in acquiring information and transmitting information up the lender's organizational hierarchy should reduce the lender's informational advantage and increase the likelihood that a loan can be sold (Stein 2002). Borrower-to-lender-HQ distance is probably capturing the second friction more than the first. Ideally, I would observe the bank branch where the borrower applied for the mortgage and measure borrower-to-branch distance to capture the first friction. La Cava (2013) gets around this difficulty by calculating the distance from the borrower's property to the nearest bank branch.[2] I instead use the timing of pairwise interstate bank deregulation to construct a proxy for borrower-to-branch distance. Prior to 1995, commercial banks were subject to interstate banking restrictions that were determined on a state-by-state basis.[3] Interstate bank laws did not prohibit lending across state lines, only the ownership of deposit-taking branches. So for the years 1990-94, I observe whether the bank holding company (BHC) could have legally owned a bank operating branches in the borrower's home state. I find that BHCs' private mortgage sale rates were significantly higher in states where they could not legally operate a deposit-taking bank branch than in states where they could. This is true even after controlling for lender×year fixed effects and the distance from the BHC headquarters to the borrower. The difference between deregulated and non-deregulated states is only present above the CLL, where lenders have greater incentives to acquire information about borrowers. This is still further evidence that the informational advantage of the lender strongly affects the selection of mortgages that are offered for sale.

In the next section I set up and solve a model of strategic interaction between mortgage lenders and secondary market purchasers. In Section 3 I discuss the empirical methodology and construction of my dataset. In Section 4 I present baseline results for the 1990s. In Section 5 I discuss the impact of Fannie and Freddie on information asymmetries in the mortgage market using a setup similar to a regression discontinuity design. In Section 6 I present several extensions to my results. First I calculate an alternate measure of distance based on the timing of interstate bank deregulation, and second I extend my sample period to the 2000s. Section 7 concludes.

2.2 Model

In this section I present a dynamic, one-shot game of incomplete and imperfect information between two players: a Bank and a Purchaser. The model incorporates a hidden-action, agency problem between the Bank and the Purchaser – the Purchaser would like the Bank to exert costly effort to audit borrowers and originate high-quality loans – and a hidden-information, adverse selection problem – if the Bank

[2]Branch location data is published in the FDIC's Summary of Deposits beginning in 1994.

[3]I use the terms "lender" and "bank" interchangeably in this paper, except for when the distinction is crucial, as it is when discussing interstate bank deregulation.

exerts origination effort, it also observes a private signal about expected borrower payoffs. The key feature of this information structure is its *non-separability*: the Bank always obtains private information when it exerts origination effort.[4] Auditing costs and benefits are both increasing with lender-borrower distance. The precision of the Bank's private signal remains constant, but due to the non-separability of auditing and learning, the cost of acquiring the signal is increasing with distance. As in Frankel and Jin (2011), the Bank acquires less information about distant borrowers than about nearby borrowers, although in that paper it is an assumption whereas in this (admittedly much simpler) model it is an equilibrium outcome. The main equilibrium is a mixed strategy: the Bank audits some of the time so as to keep the Purchaser indifferent between buying and not buying, selling a mixture of lemon loans that it privately knows to be below-average and unaudited loans, while the Purchaser buys some of the time so as to keep the Bank indifferent between auditing and not auditing. If auditing costs are bigger than benefits, the lemon effect dominates and the Purchaser's equilibrium buy probability increases with lender-borrower distance.

Both players are risk-neutral, so the payoffs at all terminal nodes are just the expected values. The full setup is as follows.

2.2.1 Setup

The model has three periods, 0, 1, and 2, and two players, a Bank and a Purchaser. A borrower is selected from the population by nature and is non-strategic.[5,6] At time 0, the Bank sequentially decides whether to audit the potential borrower, whether to originate or deny the loan, and whether to offer the loan for sale or hold it on balance sheet. At time 1, the Purchaser decides whether to buy the mortgage if offered. At time 2 the borrower's income is realized and all payoffs are made. The Bank and the Purchaser are randomly matched, so the game is one-shot, and there is no explicit allowance for reputational concerns. The reader may, however, interpret mixed strategies in terms of players returning the market and playing the game repeatedly. If the Bank tries to sell and the Purchaser does not buy, neither party may seek a second match. Funds not invested in a mortgage earn a risk-free rate of return which is normalized to 0. Play proceeds in a relatively short length of time, so there is no discounting between periods.

Play begins with the Bank's audit decision. Borrowers have two orthogonal risk dimensions: credit scores, which are publicly observable, and quality, which is initially unobservable. "Quality" refers to the aspects of a borrower's ability to repay not captured by her credit score, such as the riskiness of her income. Specifically, I suppose that some borrowers will suffer negative income shocks and not be able

[4]This feature is inspired by Vanasco (2013).

[5]For ease of exposition, I will refer to players using the gender-neutral pronoun "it" and borrowers as "she."

[6]I discuss strategic interpretations of her motivation later.

to repay the mortgage at time 2. The Bank has access to an auditing technology which enables it to draw a high-quality, creditworthy borrower (who can repay fully in expectation) at cost $\phi > 0$. The interpretation is that auditing shifts the entire distribution of borrowers to the right, such as if the Bank invests more resources in hiring and training loan officers. Other interpretations are that the Bank can choose how much to spend on advertising to attract high-quality borrowers, or that it faces agency costs with mortgage brokers and must exert effort to monitor them, or that it chooses whether or not to offer financial advice to first-time borrowers. The auditing decision is made before nature selects a borrower and is unobservable to the Purchaser.

Nature then selects a borrower with a publicly-known credit score m to apply for a mortgage from the Bank. The borrower's credit score represents her true probability of repayment: with probability m she attempt to repay the loan in good faith, while with probability $1-m$ she will default on the loan and pay nothing. The borrower has no initial endowment but wishes to purchase a house costing 1 unit of capital. The Bank has an initial endowment of funds sufficient to extend the loan and faces no capital constraints. For simplicity I assume that the house has no value as collateral in default states. The mortgage carries an interest factor $R \geq 1$ which depends on the borrower's credit score and is determined by a zero-profits condition in competitive capital markets: $mR = 1$.[7] High quality borrowers are creditworthy and will earn or save enough income to fully repay the mortgage, while low quality are not creditworthy and will only repay $Q < R$:

$$\begin{aligned} & m\left[P(\text{creditworthy})R + P(\text{not creditworthy}))Q\right] + (1-m)\cdot 0 \\ & \quad = mR - m\left[P(\text{not creditworthy})(R-Q)\right] \\ & \quad = mR - \kappa \end{aligned}$$

If the Bank audits, then it draws a creditworthy borrower and the mortgage provides a payoff of mR in expectation. If the Bank chose not to audit, then the mortgage's expected cash flows are reduced by $\kappa > 0$. This term reflects a sort of "moral hazard penalty" if the Bank shirks from acquiring information about borrower quality.

During the audit process, the Bank observes a private signal which leads it to update its beliefs about the borrower's repayment probability. With probability g the Bank learns that the borrower is a "good" type who will repay the loan with a higher-than-expected probability $p_2 > m$, while with probability $1-g$ it learns that she is a "bad " type who will repay the loan with a lower-than-expected probability $p_1 < m$. The Bank may not price on its private signal, since interest rates are set externally and depend only on m.

After deciding whether or not to audit, the Bank must decide whether to originate or deny the loan application, and whether to hold or attempt to sell the mortgage

[7] This break-even assumption is not critical, but greatly simplifies the math. What is critical is that the interest rate may only depend on the borrower's publicly-observable credit score m.

if it originates. If it originates and attempts to sell, play proceeds to time 1. A randomly-matched Purchaser is given the opportunity to buy the mortgage. The Purchaser knows the borrower's credit score m. It does not know whether or not the Bank has chosen to audit, so it does not know whether her income is sufficient to repay the loan, and it does not know the borrower's type (p_1 or p_2). Information is both imperfect (the Bank's action is hidden) and incomplete (the Bank has access to private information about payoffs). This situation is depicted in Figure 2.1: there are three nodes in the Purchaser's information set if the game proceeds to a point where it gets to play.

Following Glaeser and Kallal (1997), I introduce a parameter to allow for gains to trade: if a successful sale occurs, surplus $\theta > 0$ is created and may be divided between the two parties. The actual division of this surplus occurs via a separate bargaining game which I leave unspecified; the equilibrium agreement distributes a fraction β to the Bank and $1 - \beta$ to the Purchaser. Both parties solve backwards from their beliefs about the eventual purchase price, which in equilibrium are the same.[8] In case of a sale, the Bank's payoff is the purchase price of the mortgage, $1 + \beta\theta$, less any auditing costs it incurred, while the Purchaser receives cash flows $p_\tau R + (1 - \beta)\theta$ depending on whether the borrower is type 1 or type 2. If the Bank did not audit, the cash flow is $p_2 R - \kappa$ if the borrower is good and $p_1 R - \kappa$ if the borrower is bad.

I complete the description of the game by making the following assumptions about the values of various parameters.

Assumption 1. Credit scores are unbiased: $m = gp_2 + (1 - g)p_1$

Assumption 2. Good types are very creditworthy: $p_2 R > 1 + \theta$.

Assumption 3. Lemons condition: $p_1 R + \theta < 1$

Assumption 4. Gains-to-trade conditions.

- **(a)** Existence: $\theta > \kappa$
- **(b)** Bargaining preserves the gains to trade: $\beta \in (\frac{\kappa}{\theta}, 1 - \frac{\kappa}{\theta})$.

Assumption 5. Not too many bad types: $g > \frac{1-\beta}{2-\beta}$.

Assumption 1 states that credit scores reflect borrowers' true repayment probabilities on average. Assumption 2 states that holding a good type loan to maturity is strictly preferred to selling. Assumption 3 states that the outside option – doing nothing and earning a payoff of 1 – is strictly preferred to holding a bad type loan to maturity, even after adding in the entire gains-to-trade term. Assumption 4(a) states that the gains to trade are large enough to sustain an active secondary market if the Bank

[8]The price could also be pinned down by specifying the Purchaser's cost of raising capital.

does not audit. Assumption 4(b) guarantees that both parties receive a "fair share" of the trade surplus.[9] Assumption 5 states that the market contains a mixture of reasonably creditworthy borrowers (for example, $g > 1/2$).

Figure 2.1 shows the extensive form of the game, summarizing the decision structure and payoffs I have just described. The Bank's payoffs are listed first and the Purchaser's payoffs second.

2.2.2 Distance and Asymmetric Information

The trade-off between lender-borrower and lender-purchaser asymmetric information is the central tension in this paper. On the one hand, the Purchaser would like the Bank to become more informed and search for creditworthy borrowers. On the other hand, the Purchaser would not like the Bank to become too well informed; otherwise it is likely to face a lemons problem.

The literature suggests that the first asymmetry is increasing, and the second decreasing, as lender-borrower distance increases. Communication between distant lenders and borrowers is more formal, less frequent, and it is more difficult to transmit soft information between the parties (Berger et al. 2005). More distant borrowers face lower search costs in switching lenders (Degryse and Ongena 2005) so may be less likely to have an existing relationship with the bank (Petersen and Rajan 1994, Berger and Udell 1995). This puts lenders and purchasers equally at a disadvantage vis-à-vis borrowers. A simple way to capture these two effects is with linear specifications:

$$\phi = \phi_0 + c_\phi d \tag{2.1}$$

$$\kappa = \kappa_0 + c_\kappa d \tag{2.2}$$

where d denotes distance between the borrower and the Bank, with $\phi_0 \geq 0$, $\kappa_0 \geq 0$, and $c_\phi \geq 0$. Equation 2.1 states that the effort of searching for and originating loans to creditworthy borrowers is increasing with lender-borrower distance: ϕ is weakly increasing with d. Since the Bank is less likely to exert effort, and consequently to acquire private information about the borrower, the asymmetry of information between the Bank and the Purchaser is decreasing with distance. Equation 2.2 allows for the quality of the pool of potential borrowers to deteriorate with distance when the Bank does not audit. This could be because the quality of the Bank's prior information set (when it does not audit) deteriorates with distance due to less local knowledge (Winton 1999). Alternately, strategic borrowers who have private information about their own future cash flows might choose to apply for loan applications with more distant Banks, which face a higher cost of acquiring information about them. Empirically, c_κ appears to be positive (La Cava 2013), but this sign restriction is not necessary for the main properties of the equilibrium to hold.[10]

[9] Assumption 4(a) is actually redundant, since 4(b) implies that $\theta > 2\kappa$.

[10] A negative coefficient would reinforce rather than confound the lemons channel.

2.2.3 Equilibrium Strategies

Since the game tree has no subgames other than the entire game, I need to invoke a refinement of subgame perfect Nash equilibrium which applies sequential rationality to non-singleton decision nodes. I use the notion of sequential equilibrium proposed by Kreps and Wilson (1982).

Throughout, I will use the following notation for the behavioral strategies of each player:

- a is the probability that the Bank audits and draws only creditworthy borrowers;
- o is the probability that the Bank will originate an unknown, mixed-type mortgage (i.e., after not searching);
- o_1, o_2 are the probabilities that the Bank will knowingly originate a (bad) type 1 or (good) type 2 borrower, respectively;
- s, s_1, s_2 are the probabilities that the Bank will attempt to sell a mixed-type, type 1, or type 2 borrower, respectively;
- b is the probability that the Purchaser will buy a mortgage.

The main features of interest in the players' strategy profiles are a and b, the probabilities that the Bank audits and that the Purchaser buys, respectively. The letter *"a"* also represents how much borrower information the Bank "acquires." All proofs are provided in Appendix B.

I begin with the following preliminary results.

Lemma 2.1. *The Bank will (almost always) play pure strategies over holding and selling mortgages. Specifically:* $s^* = 1$, $s_1^* = 1$, $s_2^* = 0$.

Conditional on originating, attempting to sell is costless. If there is any positive probability that the Purchaser will buy, offering unaudited and type 1 loans for sale strictly dominates holding them, while holding type 2 loans strictly dominates offering them for sale.

Lemma 2.2. $o_2^* = 1$.

Once the auditing cost is sunk, the Bank will always originate borrowers who it has learned are the "good" type 2 kind.

Lemma 2.3. *Equilibrium origination strategies.*

(a) *If the Bank believes that the Purchaser is playing* $0 \leq b < \frac{\kappa}{\beta\theta+\kappa}$, *it will play* $o = 0$. *At* $b = \frac{\kappa}{\beta\theta+\kappa}$, *the Bank is indifferent among all strategies differing only in* $o \in [0, 1]$. *For larger values of* b, *the Bank will play* $o = 1$.

(b) *If the Bank believes that the Purchaser is playing* $0 \leq b < \frac{1-p_1R}{\beta\theta+1-p_1R}$*, it will play* $o_1 = 0$*. At* $b = \frac{1-p_1R}{\beta\theta+p_1R}$*, the Bank is indifferent among all strategies differing only in* $o_1 \in [0,1]$*. For larger values of* b*, the Bank will play* $o_1 = 1$*.*

The Bank will only originate unaudited and type 1 loans if it believes that it has a high probability of selling them. A major implication of this Lemma is the following:

Corollary 2.1. *Banks are not tempted to sell only lemons. That is, if the Bank holds beliefs such that it is willing to knowingly originate "bad" type 1 borrowers, then it is also willing to originate unaudited, mixed-type borrowers, because*

$$\frac{\kappa}{\beta\theta+\kappa} < \frac{1-p_1R}{\beta\theta+1-p_1R}$$

Intuitively, if the gains to trade θ are large enough to offset the moral hazard penalty of not auditing κ, then a mixed-type market can exist under more limited Purchaser participation than can a market with lemons. This is a statement about the Bank's incentives: the Bank will not find it optimal to supply only lemons to the market.

The following theorem gives the central result of the model.

Theorem 2.1. *Sequential equilibria with an active secondary market exist and are characterized by the following behavioral (mixed) strategies:*

$$a^* = \frac{(1-\beta)\theta - \kappa}{g(p_2R - 1 + (1-\beta)\theta) - \kappa} \tag{2.3}$$

$$b^* = \frac{\phi - \kappa}{g(p_2R - 1 - \beta\theta) - \kappa} \tag{2.4}$$

The remainder of the Bank's strategy is dictated by Lemmas 2.1, 2.2, and 2.3.: the Bank originates all types, holds good types, and sells bad and mixed types.

Parameters must be consistent with $0 \leq a^* \leq 1$*,* $\frac{1-p_1R}{\beta\theta+1-p_1R} \leq b^* \leq 1$*, and one of the two following conditions:*

i. $\kappa \leq \phi < g(p_2R - 1 - \beta\theta)$*; or*

ii. $\kappa \geq \phi > g(p_2R - 1 - \beta\theta)$*.*

Equation 2.3 gives the amount of Bank information acquisition which leaves the Purchaser indifferent between buying and not buying – i.e., the Purchaser is earning zero profits. Equation 2.4 expresses the probability of buying which leaves the Bank indifferent between auditing and not auditing. In this equilibrium, the Bank is offering both lemons and unaudited loans for sale. As a^* increases, the relative fraction of lemons in the secondary market increases. The Purchaser is losing money

on the lemons but earning positive profits on the unaudited loans. It is the cross-subsidization between lemons and unaudited loans which allows the secondary market to function.

Existence of this equilibrium rests critically on Assumption 4, that $\kappa \leq (1-\beta)\theta$. Were κ any larger, the Purchaser's share of gains to trade would not offset its expected loss from purchasing an unaudited loan. Secondary markets would break down.

Other, pure-strategy equilibria exist. I discuss these in Appendix A.

Distance and the Mortgage Sale Rate

From the discussion in Section 2.2.2, we have the following results.

Corollary 2.2. *Distance and equilibrium behavioral strategies.*

(a) *a^* is a function of lender-borrower distance* d*, with*

$$\frac{\partial a^*}{\partial d} = -c_\kappa \cdot \frac{g[p_2R - 1 + (1-\beta)\theta] - (1-\beta)\theta}{\{g[p_2R - 1 + (1-\beta)\theta] - \kappa\}^2}$$

The direction of the relationship is equal to $-1 \times \text{sign}(c_\kappa)$.

(b) *b^* is a function of lender-borrower distance* d*, with*

$$\frac{\partial b^*}{\partial d} = \frac{[g(p_2R - 1 - \beta\theta) - \kappa](c_\phi - c_\kappa) + (\phi - \kappa)c_\kappa}{\{g(p_2R - 1 - \beta\theta) - \kappa\}^2}$$

Let $\delta := \max_{d \leq (1-\beta)\theta} \left(\frac{g(p_2R-1-\beta\theta)-\phi}{g(p_2R-1-\beta\theta)-\kappa} \right) < 1$*, and suppose that* $-c_\phi < c_\kappa < c_\phi/\delta$*. Then*

i. *If* $g(p_2R - 1 - \beta\theta > \phi > \kappa$*, then the sign of the partial derivative is positive.*

ii. *If* $\kappa > \phi > g(p_2R - 1 - \beta\theta)$*, then the sign of the partial derivative is negative.*

In either case, if $c_\phi = c_\kappa = 0$*, then the partial derivative equals zero.*

Part (a) states that the auditing rate will decrease with distance if the benefit from auditing is increasing. This is because the Purchaser's intensive profit margin is damaged, and the Bank must deliver fewer lemons to keep the Purchaser in the market. Part (b) states that if the benefits of auditing are not increasing with distance too quickly (e.g., $c_\kappa \approx c_\phi$), then the buy rate will increase with distance if costs exceed benefits and decrease if benefits exceed costs. When auditing costs are large, lender-purchaser asymmetries of information are more important and the lemon effect dominates. Greater lender-borrower distance is signaling that the Bank acquired less information, so there are fewer lemons on the market and the Purchaser benefits. When auditing costs are small, lender-borrower asymmetries are more

important and the quality effect dominates. Greater lender-borrower distance is signaling lower borrower quality, which is to the Purchaser's disadvantage.

An econometrician studying the home mortgage market cannot directly observe players' mixed strategies to acquire information and buy mortgages being offered for sale. He or she will only observe equilibrium quantity (and possibly price) data on the volume of mortgages originated and the volume sold. However, this information may be related back to Bank and Purchaser strategies:

$$SaleRate = \frac{\text{Volume Sold}}{\text{Volume Originated}} = \frac{[1 - ag] \cdot b}{1} \tag{2.5}$$

In the secondary market equilibrium described by Theorem 2.1, 100% of mortgages are originated (so the denominator is equal to unity), but the Bank audits a fraction a and retains a fraction g of the audited loans.

The coefficient in a regression of mortgage sale rates on lender-borrower distance is essentially the partial derivative of 2.5 with respect to distance:

$$\frac{\partial SaleRate}{\partial d} = (1 - ag)\frac{\partial b}{\partial d} - \frac{\partial a}{\partial d}gb \tag{2.6}$$

Maintaining the conditions of Corollary 2.2 allows us to sign this partial derivative.

1. $\phi > \kappa$. Auditing costs outweigh benefits, so the lemons channel is stronger than the quality channel. The coefficient on distance is positive.

2. $\kappa > \phi$. Auditing benefits outweigh costs, so the quality channel is stronger than the lemons channel. The coefficient on distance is negative for small values of distance, but may turn positive for large values of distance.

In case 2, the conditions of Corollary 2.2 are sufficient that $\partial SaleRate/\partial d < 0$ for small values of d. However, the second derivative is positive, so the coefficient could switch signs as distance increases. If c_κ is sufficiently small, then the decrease in b is not offset by a decrease in a, and the coefficient on distance is always negative.[11]

2.3 Data and Methodology

2.3.1 Data

The Home Mortgage Disclosure Act (HMDA) was originally passed by Congress in 1975 to ensure that financing needs were being met and fair lending laws followed in metropolitan areas across the country. Amendments to the law in 1989 mandated the release of detailed, application-level data beginning in 1990. HMDA covers both

[11] More precisely, the necessary and sufficient condition is that $c_\kappa(\delta + g\frac{(1-\beta)\theta - \phi}{g(p_2 R - 1 + (1-\beta)\theta) - (1-\beta)\theta)}) < c_\phi$.

depository institutions (banks, thrifts, and credit unions) and non-depository institutions (mortgage companies), as long as they pass a minimum size threshold and have a branch (for banks) or office (for mortgage companies) in an MSA.[12] The definition of a "bank branch" follows the supervisory definition, excluding ATMs and non-deposit-taking loan processing offices. A mortgage company is construed to have an office in any MSA where it has a physical location or received at least five loan applications. Despite these limitations, other authors believe that the law achieves its intent of covering home mortgage lending activities in MSAs. Berkovec and Zorn (1996) estimate that covered lenders account for approximately 80% of total U.S. mortgage originations. A 2011 report by the Housing Assistance Council found that the majority of FDIC-insured lenders excluded due to the asset threshold are located in rural areas and specialize in agriculture lending.

Beginning in 1990, covered HMDA lenders must report the credit decision on every mortgage loan application they receive, the loan amount rounded to the nearest thousand dollars; borrower covariates such as sex, ethnicity or race, and income; and what type of counterparty bought the loan if it was sold *in the same calendar year*. HMDA filers were not required to report any loan pricing information until 2004, so I focus my analysis on a quantity-based measure: the mortgage sale rate. Most loan sales probably occur within three months of origination, so there is some underreporting of sales for loans originated in the fourth quarter. Unfortunately, the public version of HMDA does not provide within-year origination dates, so I cannot make any corrections for this. Respondents must report the location of the property tied to the loan at the Census tract level if they have a branch (for banks) or office (for mortgage companies) in the corresponding MSA. Large banks are required to report the location data for all loan applications.[13] Many lenders opt to report geographic data for all loans they process. However, Census tract locations are reported significantly less often than county location in the 1990s, and tract boundaries change in the middle of the decade when HMDA switches from 1980 to 1990 Census definitions, so I choose to focus on borrower location by county in this paper.[14] An excellent overview of these and other reporting issues in HMDA is provided by Avery et al. (2007).

My initial sample is the universe of HMDA data on mortgage originations between 1990 and 2000. I restrict the sample to 1-4 family, owner-occupied, home purchase loans (so I exclude second homes, refinancings, and mortgages on multi-

[12]In 2000, banks with at least $30 million in assets and mortgage companies with at least $10 million in assets or originating at least 100 home purchase loans were required to report.

[13]In 2000, if the bank was larger than $250 million or if it was part of a holding company larger than $1 billion, then the respondent had to report the locations of all properties.

[14]Between 1990 and 1995, about 90% of owner-occupied, 1-4 family home purchase loans had state and county data in the Loan Application Registers. Data availability improves dramatically after 1995, likely with the rise of computerized reporting, and increases to 99% in 2000. Only 85% of loans included Census tract data as late as 1995, although the number again improves dramatically and reaches 98% in 2000.

unit apartment buildings). I require that the property location be reported at the county level, and I restrict my sample to the lower 48 states and D.C. As per the previous discussion, I only include counties which are part of an MSA, where HMDA coverage is the most complete. Observations with state-county FIPS codes which do not exist are thrown out. Finally, I am unable to associate geographic coordinates with a small number of lenders (see below). The basic observation in my final panel dataset is the set of loans originated between a banking organization b and the set of customers whose properties are located in a county c.

To measure distance between banks and their customers, I need to associate both parties with specific geographic locations. I aggregate mortgage loans to the county level based on HMDA-reported property locations and map them to population-weighted county centroid coordinates from Census 2000.[15] The literature on soft information has pointed to hierarchical and geographic distance between local loan officers and distant supervisors as a key friction that increases the cost of information acquisition (Stein 2002, Liberti and Mian 2009). Accordingly, my primary measure of distance is the point-to-point distance between the lender's headquarters location and the borrower's property location. This incorporates two layers of asymmetric information between the household and the originator: the costly acquisition of soft information due to distance between the borrower and the loan officer, and the costly transmission of soft information due to distance between the local loan officer who implements lending policy and higher-ups at the bank headquarters who set and monitor lending policy. In an extension in Section 2.6.1, I use an alternate distance measure based on distance between the borrower and the nearest possible bank branch location to separate the effects of these two layers.

Avery et al. (2007) have produced a link file which matches HMDA respondents with the names and Federal Reserve RSSD IDs of their regulatory high holders.[16] Banks not part of a BHC are considered their own high holders. I pull banks' addresses from the filing year's December Call or Y-9C report, giving me city, county, and state information.[17] Savings banks and thrift holding companies file different reports, so I instead use the self-reported address of the principal subsidiary bank from the HMDA forms.[18] For respondents that are not part of a bank or thrift

[15]Obtained using the MABLE/Geocorr2K tool maintained by the Missouri Census Data Center.

[16]I use the 2011 version. Thanks to John Mondragon for assistance with this file.

[17]Call and Y-9C reports are downloaded from the Chicago Fed website. If the institution did not file in December of the HMDA year, I check for a filing in December of the previous year, of the next year, and of two years prior, in that order.

[18]The Avery et al. (2007) link file provide a state location for all high holders in HMDA. After cleaning the names of HMDA respondents and high holders, I search for the closest name match by Levenshtein distance among all subsidiaries which are both located in the same state as the high holder and are a Soundex name match. I am able to identify "principal banks" for 59,322 of 62,928 unmatched high holder-year observations. To validate the procedure, I also run it on institutions previously matched to a Call or Y-9C filing. I am able to identify a name and state match in 48,995 of 78,624 cases, among which the procedure identifies the correct city in 43,298 cases.

holding company, I use the self-reported address from HMDA. In a small number of cases, Avery et al. have identified multiple mortgage companies which are part of the same organization; in these cases I use the modal city-state.

After associating most HMDA lenders with a headquarters' city-county-state location, I acquire geographic latitude and longitude coordinates for each city from the Census 2000 Gazetteer files for places and county subdivisions. In cases where the place listed in a bank's address does not appear in the Gazetteer files, I use the county centroid coordinates (calculated with the **sp** GIS package in R[19] based on 1990 county cartographic boundary files published by the Census Bureau).

With geographic coordinates for bank headquarters and customers in hand, I calculate great circle, "as the crow flies" distance between the two locations using the spherical law of cosines.

HMDA does not report the identity of the purchasing institution in cases where a loan is originated and sold, but it does report whether the purchaser was private or government-owned or sponsored (including Fannie Mae, Freddie Mac, Ginnie Mae, and Farmer Mac). This allows me to calculate the total and private mortgage sale rate for the portfolio of loans originated by bank b in county c and year t. I am also able to measure each borrower's loan-to-income ratios as a proxy for creditworthiness or mortgage affordability.

I obtain additional co-variates based on borrower geography: annual county median household income from the Census Bureau Small Area Income and Poverty Estimates program, county population density from the 2000 Census, and annual state house price indices from the FHFA all-transactions series (average of quarterly, not-seasonally adjusted).[20]

Since the purchasing power implied by nominal household income is specific to the time and place where that household resides, it is useful to remap these income figures into percentile rankings which may be more easily compared. I adopt a parametric approach by assuming that the income distribution in each county follows a two-parameter gamma distribution. The SAIPE program reports median household income and the percent of people in poverty for U.S. counties since 1989. For each county-year, this produces a system of two nonlinear equations in two unknown parameters. I solve the system numerically and find that for the vast majority of county-years, either one or two solutions exist. If two solutions exist and one of them has a shape parameter $\alpha > 1$, I use this solution. Otherwise I use the solution with the largest value of $\alpha < 1$. When no solution exists, I choose an approximate solution by either top- or bottom-coding the value of α from other counties in that year, then locate the value of β solving that county-year's median income equation.[21]

[19] Pebesema and Bivand (2005), version 1.0.9.

[20] SAIPE was not annual until 1997. I use log-linear interpolation to create an annual series for median HH income.

[21] For example, non-existence occasionally occurs because the percent of people in poverty is greater than 0.5 but the poverty threshold is smaller than county median income. The SAIPE figures

The procedure is described in detail in Appendix C.

2.3.2 Methodology

Identifying Assumptions

The discussion in Section 2.2 has motivated the following style of regression:

$$SaleRate_{bct} = \alpha_t + \alpha_b + \beta d_{bc} + \gamma' x_{bct} + \varepsilon_{bct} \tag{2.7}$$

The dependent variable is the sale rate observed by the econometrician between bank b and the group of customers c taking out mortgages on properties in the same location. The parameter of interest is the coefficient on the distance d_{bc} between bank b and customers c. The model I have discussed pertains to customers with similar observable risk characteristics applying for similar-sized mortgages from the same type of bank. It will thus be natural to include control variables x_{bct} relating to observable characteristics of the bank, the customers, and the housing market in their shared geography. Time fixed effects will absorb interest rate conditions that vary from year to year and affect all MSAs, while bank fixed effects can hold the lender constant.

The model allows for endogenous customers and market scope. On the first point, customers may strategically choose their banks – so in particular, unobservable customer quality is allowed to covary with bank-customer distance. On the second point, I endogenize the origination decision, so banks may respond rationally to possibly higher screening costs of more distant borrowers by exiting the market. However, it will be important to assume that the purchaser's buy decision is exogenous to lender-borrower distance except through the signals distance conveys about bank private information and loan quality. If this assumption holds, OLS will provide a consistent estimate of the parameter of interest, which is simply the population best linear predictor coefficient.

This assumption could be violated in several ways. First, suppose that lenders with different geographic scopes are also heterogeneous in their participation in secondary markets. For example, it may be that large banks which do more distant lending also have easier access to secondary markets. I attempt to control for lender heterogeneity via size and geographic scope controls in some specifications, and via lender×year fixed effects in other specifications. Second, suppose that purchasers wish to hold a geographically diversified portfolio of loans, so secondary-market demand is systematically higher in markets with fewer banks (where average lender-borrower distance is also higher). I will attempt to address this concern by controlling for market concentration via a Herfindahl-Hirschman index.

---

are estimates, referring to *households* (for median income) and *people* (for percent in poverty), while the poverty threshold refers to *families*, so some inconsistencies are to be expected.

Predictions

According to the model presented in Section 2.2, the coefficient on distance will always be positive if the lemons channel is stronger than the quality channel, while it will be at least sometimes negative if the reverse is true.

Prediction 1. If the severity of asymmetric information between lenders and *purchasers* is greater than the severity of asymmetric information between lenders and *borrowers*, then the coefficient on distance $\beta > 0$. If the severity of asymmetric information between lenders and borrowers is greater than the severity of asymmetric information between lenders and purchasers, then $\beta < 0$.

To help pin down that any distance effect I find is driven by informational asymmetries, and not some other mechanism, I exploit a discontinuity around the national conforming loan limit. Above this hard dollar limit, the GSEs are not allowed to purchase home mortgage loans, so all mortgage sales in the "jumbo" market are to private counterparties. Below the limit, the GSEs post guidelines based on borrower credit scores, total debt-to-income ratios, the size of mortgage downpayments, and whether or not the borrower purchases mortgage insurance. Mortgages meeting these guidelines are considered "conforming" and are eligible for delivery to Fannie and Freddie. There is a sharp drop in secondary market liquidity at the conforming loan limit (Figure 2.3). As such, seller incentives to acquire information about borrowers rise sharply at the CLL, since there is a much lower probability they will be able to successfully sell a jumbo loan.

My empirical strategy is to use the discontinuity in market liquidity due to the legal restriction on GSE purchases around the conforming loan limit to explore the impact on informational asymmetries in the *private* segment of the secondary market. Consider the following RDD:

$$\begin{aligned} PvtSaleRate^{J}_{bct} - PvtSaleRate^{C}_{bct} &= (\alpha^{J}_{t} - \alpha^{C}_{t}) + (\alpha^{J}_{b} - \alpha^{C}_{t}) + (\beta^{J} - \beta^{C})d_{bc} \\ &+ (\gamma^{J} - \gamma^{C})' x_{bct} + \nu_{bct} \end{aligned} \tag{2.8}$$

Prediction 2. If the lemons effect dominates the quality effect, then the *difference* in coefficients on distance, $\beta^{J} - \beta^{C}$, will be positive. If the quality effect dominates the lemons effect, then the difference in coefficients $\beta^{J} - \beta^{C}$ will be negative.

The style of test proposed in this paper allows us to assess whether distance is really picking up informational asymmetries and helps rule out alternate explanations such as foreclosure costs. This test does not let us make normative statements about the role of Fannie and Freddie on secondary mortgage markets. This is because loan size is a choice variable, so there is great potential for borrowers and lenders to sort around the CLL and violate the internal validity requirements necessary for an RDD

to consistently estimate a local average treatment effect (Hahn et al. 2001). I will present evidence that such borrower sorting does not contaminate my estimates of a discontinuity in the coefficient on distance, $\beta^J - \beta^C$, below.

2.4 Baseline Results

I begin by estimating equation 2.7 using the total mortgage sale rate as the dependent variable. The sample period is 1990-2000. Each observation represents the set of loans that a bank b makes to customers living in the same county c, where I restrict to counties in MSAs in the lower 48 states and D.C. Results are presented in Table 2.1.

In column 1 I include controls based on common borrower geography – log county population density, log median household income in 2000 and its growth rate over 1990-2000, and annual house price appreciation at the state level – and lender size and geographic scope. "Lenders" are aggregated to the regulatory high-holder (bank holding company) level. All standard errors are clustered by lender.

The estimated coefficients indicate mortgages originated in high-income and urban counties are easier to sell. For example, a doubling of median household income in 2000 is associated with a 9.5 percentage point increase in the total mortgage sale rate. The effect of state house price appreciation is insignificant at standard levels. Measuring lender size by the total number of loans originated in HMDA for that year, larger lenders sell more mortgages on secondary markets (a doubling of lender size is associated with a 5.7 percentage point increase in the total mortgage sale rate). Controlling for lender size, geographic scope (the number of states in which the lender originates at least 1 loan) is negatively associated with mortgage sale rates. Both effects are significant at smaller than a 1% level. Lenders who originate a greater fraction of loans above the CLL also have a lower overall sale rate, which makes sense because Fannie and Freddie are not allowed to purchase loans in this market segment. Finally, I note that the coefficient on the county "competition index" ($= 1 - HHI_{ct}$) is small, negative, and insignificant. The negative sign of the estimate is consistent with a purchaser-diversification story (fewer lenders in a county are associated with less competition and higher sale rates), but the evidence is extremely weak.

The main coefficient of interest is the effect of distance on mortgage sale rates. I estimate that a doubling of lender-HQ-to-borrower-county-centroid distance is associated with a 2.5 percentage point increase in the total mortgage sale rate. Put another way, the standard deviation in the log of distance is 1.84, while the standard deviation in log median household income is 0.239, so a one-S.D. increase in the log of distance has the same impact on mortgage sale rates as a two-S.D. increase in the log of county median household income in 2000.

It is possible that these findings reflect not the increasing cost of bank information

acquisition, but the fact that different lenders have different business models which are correlated both with borrower distance and with mortgage sale rates. Were this the case, distance should lose its explanatory power once I include lender fixed effects. Table 2.1, column 2 reports the results of this specification. Within the same lender, HQ-to-borrower distance has essentially no explanatory power on the total mortgage sale rate. This is consistent with informational matching a la Stein (2002) and Berger et al. (2005), but it is also consistent with other stories of lender-borrower matching which have nothing to do with informational asymmetries. For example, if the partial derivatives of ϕ and κ with respect to distance both equal zero – the cost of acquiring information and the benefit from acquiring information do not depend on distance – then Corollary 2.2 states that the coefficient on distance will equal zero.

A third possibility is that informational asymmetries are not very important for the average mortgage borrower. If the average mortgage borrower in the 1990s is basically creditworthy (in terms of both credit score and income), then private bank information might simply be unimportant. A strategy to test this hypothesis is to re-estimate equation 2.7 for low-creditworthiness subgroups. In particular, consider mortgage borrowers in the bottom of their state's income distribution. These are marginal borrowers –they are most likely younger borrowers, they may have shorter credit histories, many of them may be first-time homeowners, and they are probably on the margin between buying and renting. Private bank information should be particularly relevant for this group of borrowers.

Columns 3 and 4 of Table 2.1 re-estimate equation 2.7 for borrowers in the bottom quintile of their state's income distribution. The 20th-percentile threshold in each state-year is estimated using the procedure described in Appendix C, based on state estimates of median household income and fraction of people below the poverty line. I estimate that a doubling of lender-borrower distance is associated with a 2.26 percentage point increase in the mortgage sale rate; the effect is significant at the 1% level. This is very close in magnitude to the estimate from column 1, for the entire population of HMDA borrowers. Moreover, the effect remains similar in magnitude and highly significant when I include lender and lender×year fixed effects in column 4. Within the same lender, low-income-borrower loans are more likely to be sold if the borrower resides farther away from the lender's headquarters. This is consistent with a lemon effect dominating a quality effect: purchasers prefer lenders to be uninformed about borrower quality.

Columns 5 and 6 compare the mortgage sale rate for lenders in the bottom versus the top quintile of each state's income distribution. The dependent variable is the difference $SaleRate_{bct}^{Q1} - SaleRate_{bct}^{Q5}$, so the sample only includes lenders originating loans to borrowers in both income quintiles in the same county-year. Column 5 reports results without lender fixed effects and column 6 with lender and lender-year fixed effects. The finding is that greater lender-borrower distance is *differentially* associated with higher mortgage sale rates for the lowest versus the highest income

borrowers. Since private bank information is most likely to matter for low income borrowers, this is consistent with the interpretation that distance is picking up the cost of bank information acquisition rather than something else. I estimate that a doubling of distance is associated with a 0.6 to 1.1 percentage point differential increase in the bottom-quintile mortgage sale rate (both significant at the 1% level).

To summarize, my baseline results show that lender-borrower distance has a powerful predictive effect on the total mortgage sale rate. For low income borrowers, about whom private information is likely to be informative, the effect is robust to lender fixed effects. The effect of distance is differentially larger for low income versus high income borrowers, suggesting that distance is indeed picking up an increasing bank cost to acquire borrower information. Finally, the estimated coefficient is positive and similar in magnitude across five of six specifications. This indicates that the lemons channel dominates the quality channel. It is easier for lenders to sell loans on secondary markets when they are *less* informed about the borrower, not *more* informed.

2.5 The Effect of GSEs on Information Asymmetry

While I interpret the previous positive coefficient on distance in terms of asymmetric information – demand is increasing with distance, because purchasers believe distant loans are less likely to be lemons – there are other possible explanations. One leading explanation is that lenders face higher costs in managing the foreclosures of more distant properties, so the supply of loans being offered for sale is increasing with distance. To help pin down the link between distance and asymmetric information between lenders and purchasers, I turn to the subsample of loans made in the vicinity of the conforming loan limit. Specifically, I focus on mortgage loans with principal balances (at time of origination) between $[CLL - \$10,000,\ CLL + \$10,000]$, and estimate equation 2.8. The dependent variable is the *private* mortgage sale rate, defined as the number of loans sold to private counterparties divided by the number of loans originated in a bank-county-year portfolio. I predict that the effect of distance on a mortgage's probability of being sold should be larger for a jumbo mortgage than for an otherwise-comparable conforming mortgage.

2.5.1 Assignment, Sorting, and Internal Validity

Figure 2.2 shows that the assignment rule is binding. The probability of a loan being "treated" – i.e., sold to a GSE or wholly-owned government corporation – declines from around 50% just below the CLL to around 8% just above the CLL. The decline in treatment probability is less than 100% for a number of reasons. First, loan size is one of several dimensions which jointly determine whether a mortgage

is conforming and thus eligible for delivery to Fannie and Freddie, so the design is "fuzzy" rather than "sharp" in the RDD lingo. Second, treatment is not a deterministic function of assignment because banks may choose not to sell some conforming loans. Third, the CLL does not apply to loan sales to Ginnie Mae and Farmer Mac. Fourth, the CLL is updated in October of every year, so some loans taken out just above CLL may be conforming loans that were originated and sold to Fannie and Freddie during the fourth quarter. As mentioned previously, the public version of HMDA does not provide within-year origination dates. Finally, loan amounts are rounded to the nearest thousand dollars in HMDA, so the "zero" cell contains loans both below and above the CLL. Since it is not obvious how to classify the zero cell, I will exclude it from all analyses.

Figure 2.3 shows the average volume of 1-4 family, owner-occupied home purchase loans originated each year between 1990 and 2000. The sharp drop in loan volume just above the conforming loan limit is consistent with a reduction in mortgage supply in the jumbo-size market. Origination volume declines from around 18,000 per year for loans $1,000 below the CLL to approximately 1,000 per year for loans $1,000 above the CLL. The GSEs appear to facilitate an increase in the supply of mortgage credit in the conforming-size market. Authors using other other datasets have confirmed that mortgage interest rates are lower below the CLL. However, loan size is a choice variable, so this discontinuity in the density could be evidence of borrower sorting. Sorting could cause discontinuities in borrower attributes at the CLL, which might invalidate the design (McCrary, 2008). Consider a high-credit score, low debt-to-income ratio borrower who is considering taking out a mortgage with principal CLL + $1,000, a jumbo mortgage with an accordingly high interest rate. She has a strong incentive to increase her down payment by $1,001 in order to bring the principal below the CLL and take advantage of the lower interest rate on conforming mortgages. The borrowers most likely to select into CLL - $1 mortgages are those who would otherwise qualify for a conforming loan. Moreover, they must be able to afford the additional down payment, either through their own income and savings, or from family members with the resources to lend them the additional down payment. All of these factors are likely to lead to a discontinuity in borrower attributes at the CLL: borrowers just below should have higher credit scores, more financial resources, and less overall debt than borrowers just above.

Table 2.2 looks for evidence of borrower sorting by testing for a discontinuity in observable borrower attributes at the CLL. I estimate the following regression:

$$x_{bct} = \alpha_t + m(LoanSize_{bct}) + \delta \mathbb{I}\{LoanSize_{bct} > CLL_t\} + \varepsilon_{bct} \tag{2.9}$$

Each observation is the set of borrowers living in county c taking out loans from lender b in year t, between 1990 and 2000. The function $m(LoanSize)$ is a fifth-order polynomial in the assignment variable, so each attribute x may vary in a flexible manner with loan size. I test for the presence of discontinuities in attributes by county demographics in Panel A, by borrower characteristics reported in HMDA

in Panel B, and by lender in Panel C. Column 1 reports $\hat{\delta}$, the estimated discontinuity in a given attribute at the CLL. Borrowers just above and just below the CLL do not differ by county median household income, but borrowers just above the CLL do tend to live in more unequal counties (as measured by the Gini concentration coefficient) and in denser counties. Borrowers just above the CLL are on a lower rung of the income ladder (by four percentage points along their county's income distribution) and they take out bigger loans as a percentage of their income (by 23%). Since loan size is the assignment variable, these last two facts indicate that borrowers with higher incomes sort into loans just below the CLL. Finally, lenders originating loans just below the CLL are no different than those just above by size, but they do operate in two more states on average.

While there is strong evidence of borrower sorting at the CLL, borrowers in the vicinity of the CLL look very similar *on average*. Column 2 re-estimates equation 2.9 without the function $m(LoanSize)$, so the estimate $\hat{\delta}$ now represents the difference in average value of each attribute after adjusting for year fixed effects. The average differences in attributes in column 2 are much smaller in magnitude than the discontinuities in attributes estimated in column 1. For example, column 1 reports that population density jumps by nearly 1600 people per square mile at the CLL, while column 2 reports that the typical borrower above the CLL lives in a county with only 740 more people per square mile on average. Despite the evidence of income-based sorting around the CLL reported in column 1, borrowers above and below earn very similar incomes on average. Although income jumps discontinuously *down* at the CLL, average incomes above the CLL are 0.3 percentage points *higher* within borrowers' respective county income distributions. Moreover, loan-to-income ratios are only 5 percent-of-income higher above the CLL on average, despite the 23 percent-of-income jump at the CLL.

Focusing on borrowers within \$10,000 of the CLL thus appears to provide a sample of borrowers with similar attributes above and below the CLL on average. Moreover, the object of interest in my analysis is not an average treatment effect, so the usual concerns about selection effects are more subtle. Borrower sorting around the cutoff will not invalidate my results as long as the relative sorting of borrowers by distance from lender stays the same. For example, it is not necessarily worrisome if borrowers just below the CLL have higher incomes than borrowers just above, but it would be worrisome if the difference in incomes for faraway versus nearby borrowers jumps discontinuously at the CLL.[22] It seems very unlikely that the geographic arrangement of borrowers changes discontinuously at the CLL. Nevertheless, I will address this concern in a series of robustness checks after presenting the main results.

[22]Econometrically, correlation between $x_{bct}$and $\mathbb{I}\{Jumbo_{bct}\}$ is not problematic, but correlation between x_{bct} and $Dist_{bc} \times \mathbb{I}\{Jumbo_{bct}\}$ is.

2.5.2 Regression Discontinuity Analysis

In section 2.4 I showed evidence consistent with lender private information hindering trade on secondary mortgage markets. If Fannie and Freddie alleviate this problem by increasing market liquidity and reducing seller incentives to acquire information, then the coefficient on distance should be more positive above the CLL, where they are barred from operating. On the other hand, if the coefficient on distance is due to higher foreclosure costs to lenders, then there should be no difference in coefficients just above versus just below the CLL.

The main results are presented graphically in Figures 2.4 and 2.5. Figure 2.4 plots the private mortgage sale rate against loan size in the vicinity of the CLL, separately for borrowers residing near the lender and borrowers residing far away from the lender. Each data point is an average across lender-county-years by thousand dollar increment above or below the CLL, over the sample period 1990-2000. Above the CLL (on the right side of the graph), the private mortgage sale rate is increasing with distance. The average sale rate for faraway borrowers is above 30%, versus 20-25% for nearby borrowers. This is consistent with a lemons channel dominating. However, for borrowers taking out loans just below the CLL, the reverse is true. The private mortgage sale rate averages around 15% for faraway borrowers versus a little above 20% for nearby borrowers. It is easier to sell the loans of nearby customers to private purchasers in the market with Fannie and Freddie. This is consistent with a quality channel dominating; the presence of Fannie and Freddie reduces bank incentives to acquire information, so private-label purchasers prefer banks to be more well-informed and differentially buy nearby loans.

Figure 2.5 shows the same data, but plots the private mortgage sale rate in each market segment (jumbo versus conforming) against lender-borrower distance on the horizontal axis. The jumbo-conforming gap in private sale rates is zero or negative for the closest borrowers, but widens as distance increases: the sale rate for loans just below the CLL decreases and the sale rate for loans just above the CLL increases. For the most distant borrowers, the gap is large and positive. As lender-borrower distance increases, it is easier to sell loans to private counterparties in the purely private secondary market (just above the CLL) but more difficult to sell loans to private counterparties in the conforming market.

I estimate the RDD (equation 2.8) and present numerical results in Table 2.3. The "zero cell" is excluded from the analysis – since loan amounts are rounded to the nearest thousand dollars in HMDA, this cell contains loans both above and below the size threshold. Column 1 reports the baseline RDD, with year fixed effects but no other controls. All three coefficients of interest are highly significant. The estimates indicate that the private sale rate is approximately 7.7 percentage points lower in the jumbo market.[23] The coefficient on distance is negative in the conforming-size

[23]Since this column does not control for observable borrower quality, the negative coefficient on $\mathbb{I}\{Jumbo_{bct}\}$ may be be biased due to a contemporaneous decline in borrower quality just above

market (-2.1) but positive in the jumbo-size market ($-2.1 + 3.1 \approx 1.0$). This is consistent with a lemons channel determining purchasers' buy strategies above the CLL, but a quality channel determining the buy strategy below the CLL. Banks prefer to sell and private counterparty purchasers prefer to buy nearby loans below the CLL but faraway loans above the CLL.

As discussed in section 2.5.1, loan size is a choice variable, so higher income borrowers who can qualify for a conforming loan have great incentives to select into smaller loans which fall just below the CLL. This creates a discontinuity in borrower attributes at the conforming loan limit. However, borrower sorting does not pose a challenge to the internal validity of my RDD unless the spatial arrangement of borrowers also jumps discontinuously at the CLL. To address this concern, I add county-level demographic variables (median household income, Gini coefficient, population density, state house price growth rate, and lender competition index) and borrower-level variables from HMDA (loan-to-income ratio, borrower income percentile ranking in county) as controls and re-estimate the RDD equation in column 2. The coefficient on distance remains negative and significant below the CLL and positive and significant above the CLL. Observable differences in borrower attributes, based on their locations and on the information reported about them in HMDA, do not seem to drive the results.[24]

Another possibility is that the observed difference in the coefficient on distance is driven by lenders with different business models operating above versus below the CLL. For example, Table 2.2 indicates that lenders operating just below the CLL have greater geographic scope – the typical lender operates in two more states than a lender just above the CLL. The coefficient on distance might be non-positive in the conforming-size market because nationwide banks engage in arms-length lending and acquire less private information about borrower quality. In column 3 I add controls for bank size, as measured by log of the number of loans originated in all MSAs nationwide, and geographic scope, measured by the number of states in which a bank originates at least one loan. The main impact of these controls is that the coefficient on distance in the conforming market falls in magnitude and is statistically indistinguishable from zero. The coefficient on log(origination volume) is negative and highly significant: for every doubling in the volume of loans a lender originates, the private sale rate declines by 3 percentage points.[25] This indicates that more distant borrowers taking out mortgages just below the CLL tend to be harder to sell because they tend to be originated by large-volume lenders who sell fewer mortgages on the private-label market.

Banks are heterogeneous for reasons beyond size and geographic scope. The

---

the CLL.

[24]Interestingly, differences in observable borrower attributes do not seem to drive the negative coefficient on $\mathbb{I}\{Jumbo_{bct}\}$, whose estimate is essentially unchanged from column 1.

[25]This differs from the results reported in Table 2.2, which indicate a positive correlation between origination volume and the total mortgage sale rate.

1990s was a period of heightened merger activity, as regulations prohibiting cross-state banking were relaxed and banking evolved from a local to a national business. Banks which specialized in information acquisition and relationship lending one year may have changed business models and specialized in volume lending and fee-based income the next (see Stiroh 2004, for example). In addition to the county and borrower controls included in columns 2 and 3, I add lender and lender×year fixed effects to my estimating equation in column 4. This controls for differences in business models which are fixed across banks and that may change within a bank over time. The comparison is now between two borrowers living in similar counties, with similar incomes and loan-to-income ratios, taking out loans from the same bank in the same year, who differ only in their distance from the bank headquarters. The estimates indicate that there is no difference in the private sale rate for conforming-size loans, but a positive and significant difference in the private sale rate for jumbo-size loans. In the market where Fannie and Freddie are prohibited from operating, a doubling in lender-borrower distance is associated with a 2.1 percentage point increase in the private mortgage sale rate. This is consistent Prediction 2 and not with a foreclosure costs story. This is not to say that foreclosure costs do not affect lenders' choice of which loans to sell, only that informational asymmetries between lenders and purchasers also seem to be a factor.

2.5.3 Robustness Checks

Global Polynomial Control Function

It is common in RDD applications to adopt a control function approach (Heckman and Robb 1985, van der Klaauw 2008) and include some function of the assignment variable as an additional control in the RDD estimating equation (equation 2.8). In this application, the assignment variable is loan size, so the goal is to capture unobserved borrower attributes that affect the private sale rate and are correlated with loan size (and thus, with whether a loan falls on the conforming or jumbo side of the cutoff). If the included function of loan size does a good job of approximating the unknown, true function, then the remaining parameters in the RDD equation may be consistently estimated.

Column 1 of Table 2.4 presents such a set of estimates. I include the same set of controls as in the last column of Table 2.3 – county demographics, borrower income percentile and LTI ratio, and lender-year fixed effects – and augment the regression by estimating the parameters of a global fifth-order polynomial in loan size. The coefficients on distance in the conforming-size and jumbo-size markets are essentially identical to those in column 4 of Table 2.3: I estimate that the private sale rate increases with distance in the latter and does not depend on distance in the former. However, the main effect of being in the jumbo market is now negative and significant, although it is less than half the size of the main effect I estimated in

Table 2.3 column 1.

The main concern in my setting is that unobserved borrower attributes may vary by distance differentially for loans above and below the CLL. If these omitted attributes are also correlated with loan size, then we would expect the coefficient on distance to vary continuously with loan size but not jump at the conforming loan limit. I test this possibility by interacting the fifth-order polynomial of loan size with distance and report the results in column 2 of Table 2.4. I do not report the main coefficient on $\log(Distance_{bc})$ in the conforming-size market, since it can properly be interpreted only in conjunction with the interacted coefficients. The coefficient of interest remains the coefficient on $\mathbb{I}\{Jumbo_{bct}\} \times \log(Distance_{bc})$, the estimated discontinuity in the effect of distance on private mortgage sale rates when Fannie and Freddie drop out of the market. This coefficient remains positive, and indeed, it is nearly identical in size and significance to previous estimates. A loan which is twice as far away from the bank is about 2 percentage points more likely to be sold if the pair are just above the CLL, differentially versus an identical pair of loans falling just below the CLL.

The analysis in section 2.5.1 suggested that borrower sorting may occur locally around the CLL, but that average observable attributes appear to be very similar. Figure 2.3 indicated a spike in the density of loan volume in the cells $[CLL - \$1,000, CLL]$. The zero cell is already excluded from analysis since it cannot be classified as either conforming or jumbo-size. As a further robustness check, I symmetrically omit cells containing loans within \$1,000 of the CLL (column 3) and \$5,000 (column 4). This allows me to exploit the discontinuity in eligibility for sales to GSEs, while excluding the sets of loans nearest the CLL which are most likely to be contaminated by borrower sorting. As before, the coefficient on distance is statistically indistinguishable from zero in the conforming-size market, but positive, significant at the 1% level, and essentially identical to previous estimates in the jumbo-size market.

Breakpoint Test

Our knowledge of the structure of secondary mortgage markets suggests that the conforming size threshold is the correct cutoff. This is confirmed by visual evidence that the GSE sale rate declines sharply (Figure 2.2) and the coefficient on distance changes signs (Figure 2.4) at the CLL. However, an alternate explanation is that the coefficient on distance varies with loan size in a random manner which cannot captured by the polynomial interaction from the previous section. Under this interpretation, there is no economic significance to the jump at the CLL: we would be likely to find a jump at any randomly-chosen breakpoint.

To consider this possibility by running a series of placebo tests with cutoffs other than the CLL. Each placebo test mimics the structure of the true RDD: I restrict the sample to a plus- and minus-\$10,000 window around the placebo cutoff, and I

omit the cell containing the cutoff. I run the placebo test for every cutoff between $CLL - \$40,000$ and $CLL + \$40,000$, for a total of 80 placebo tests. For each test I use the same specification as in column 2 of Table 2.4, controlling for borrower and county covariates and lender×year fixed effects. This specification also includes a fifth-order polynomial of loan size and its interaction with the log of distance. I use this specification for the placebo tests because by allowing the coefficient on distance to flexibly fit the data, it is very powerful at detecting local discontinuities.

The estimated coefficients on $\mathbb{I}\{Cutoff_{bct}\} \times \log(Distance_{bc})$ are plotted as bars against the placebo cutoffs in the top panel of Figure 2.6. The dashed lines around the horizontal axis show a 95% confidence interval for the estimated discontinuity centered at zero, using lender-clustered standard errors. Coefficients outside these bands are statistically different from zero at a 5% level.

Overall, we are more likely to reject the null hypothesis of no discontinuity than would be expected by chance alone. In particular, we are more likely to reject the null for samples including the CLL. For breakpoints between zero and $2,000 above the CLL, I estimate a positive discontinuity in the distance coefficient. This is consistent with previous evidence. However, when the breakpoint is moved just below, or more than $2,000 above, the CLL, the tests produce negative and significant discontinuities in the distance coefficient. Counterintuitively, this is consistent with the true breakpoint being at the CLL. When the coefficient is not allowed to jump discontinuously up at the CLL, the polynomial attempts to fit the jump continuously via a large and positive local first derivative. Placing a placebo breakpoint below the CLL allows the polynomial to jump downwards prior to reaching the CLL, then slope sharply upwards at the CLL. Placing a placebo breakpoint above the CLL allows the polynomial to avoid overshooting by jumping back downward.

The bottom panel reports the empirical distribution of t statistics over the 81 tests. These are not independent tests since the samples are dependent. Rather, the thought experiment is: if we were to randomly choose any cutoff between $CLL - \$40,000$ and $CLL + \$40,000$, what is the probability that we would observe a t statistic at least as extreme as we do in the actual test? This empirical p-value is $3/81 \approx 0.037$: only two placebo RDDs produce more extreme t statistics than the real RDD.

2.6 Extensions

2.6.1 An Alternate Measure of Lender-Borrower Distance

The results presented thus far have focused on point-to-point distance from a borrower's county population centroid to the city where the lender is headquartered. This measure of lender-borrower distance is highly correlated with hierarchical frictions within the banking organization: since soft information is costly to "harden"

and transmit, greater distance between a local loan officer and his faraway supervisor disincentivizes the loan officer from acquiring such information, easing lemon problems in secondary markets. However, borrower-to-lender-HQ distance is probably not capturing frictions between the loan officer and the borrower very well. If information acquisition costs, as opposed to information transmission costs, increase with distance, then loan officers will tend to be less informed about borrowers residing far away from the branch. We should predict that borrower-to-branch distance also alleviates the lemons problem between originators and secondary-market purchasers.

An ideal dataset would provide the geographic location of the bank branch from which each borrower took out her loan. I could then regress mortgage sale rates on borrower-to-branch distance as in Table 2.1 or 2.3. However, we might be concerned that bank branch locations are endogenous with respect to unobserved local economic conditions. The usual concern voiced by policymakers is that banks tend to locate branches in high-creditworthiness neighborhoods, so the credit needs of low-income and minority communities go unmet. But suppose the situation were reversed: in response to political or regulatory pressure, banks strategically locate more branches in geographies where borrowers are less creditworthy than expected based on observable HMDA characteristics. We would observe mortgage sale rates increasing with borrower-to-branch distance, not for informational reasons, but because borrowers near branches are systematically less creditworthy than borrowers far away from branches.

The 1990s present a neat laboratory to explore this mechanism because there is exogenous variation in bank branch location. The Douglas Amendment to the Bank Holding Company Act of 1956 gave state legislatures the discretion to allow out-of-state BHCs to acquire in-state banks (although full interstate branching remained prohibited). Prior to the late 1970s, all states prohibited interstate bank acquisitions, but during the 1980s and the early 1990s, state legislatures began relaxing these restrictions, sometimes unilaterally, sometimes on a bilateral or regional basis. The Riegle-Neal Act nationalized the process of geographic deregulation, permitting full interstate *banking* in 1995 and requiring states not opting out to allow interstate *branching* by 1997.[26]

This piece-meal approach to interstate deregulation suggests the following instrumental variables strategy. Let $Interstate_{r \to s,t} = 1$ if state s allows entry from BHCs headquartered in state r in year t, and 0 otherwise.[27] Then if we knew the identity and location of the branch with which every customer residing in state C of every BHC headquartered in state B interacted, we could implement the following two-stage procedure:

$$BranchDist_{ibct} = \xi_t + \xi_b + \eta \cdot Interstate_{B \to C,t} + \zeta' x_{bct} + \nu_{bct} \tag{2.10}$$

[26]Interstate banking laws placed restrictions on the ownership of banks and location of branches. They did not place restrictions on the locations of customers. Many banks opened non-deposit-taking loan processing offices in other states expressly for the purpose of lending across state lines.

[27]$Interstate_{r \to r,t}$ is coded as 1.

$$SaleRate_{ibct} = \alpha_t + \alpha_b + \beta \hat{BranchDist}_{ibct} + \gamma' x_{bct} + \varepsilon_{bct} \tag{2.11}$$

where $SaleRate_{ibct}$ is the set of customers residing in county c taking loans out from branch i of bank b in year t.

There are several limitations to this strategy. First, I do not know which branch a borrower actually took her loan out from.[28] Moreover, the main source of branch location data, the FDIC Summary of Deposits, is not publicly available until 1994, providing only one year of variation in interstate banking laws for the IV before Riegle-Neal went into effect and full interstate banking was allowed.[29] However, the reduced form regression of mortgage sale rates on $Interstate_{B \to C,t}$ can be estimated using five full years of data, from 1990 to 1994. Finally, the instrument is a binary variable which only measures whether or not a borrower's home state has deregulated with the BHC's home state. It neglects possibly useful geographic information about how proximate a borrower is to other states which have already deregulated.

To construct a second instrument which incorporates this information, let us define state B's "neighbors" as the set of states n_{Bt} into which BHCs headquartered in state B are allowed entry in year t. This set of relationships is summarized by a neighbor matrix, $N_t = [Interstate_{r \to s,t}]_{51 \times 51}$, wherein each element contains a one if a column state permits entry from that row state, and a zero otherwise. Since entry is not necessarily commutative, $N_t \neq N_t'$. Define d_{cs} as the distance from county c to the closest border of state s, so $d_c = [d_{cs}]_{1 \times 51}$ is a row vector containing all 51 county-to-state distances for a given county. Then the **nearest-deregulated-neighbor distance** between a bank and a county is

$$NDNDist_{bct} = \min(d_c * \iota_b' N_t) = \min_{s \in n_{Bt}} (d_{cs}) \tag{2.12}$$

where $\iota_b = [\mathbb{I}\{b\text{'s HQ in state } s\}]_{51 \times 1}$ is the column vector selecting BHC b's home state and the "$*$" operator indicates element-wise multiplication. $NDNDist_{bct}$ measures the minimum legally-permissible distance between a borrower in county c and a branch of a bank owned by BHC b. $NDNDist_{bct}$ should be correlated with the distance to the actual branch a customer uses, but it is uncorrelated with customers' strategic choices about which branch to apply to as well as bank choices about where to locate branches within a state.

I obtain data on the dates of pairwise interstate banking deregulation from Goetz et al. (2013), who use this data to analyze the impact of exogenous changes in bank geographic scope on firm valuation.[30] At year-end 1990, 1,065 of a possible $51 \cdot 50 = 2,550$ state-pair paths (41%) were open to interstate bank entry. This rose by 336 linkages to 1,401 (nearly 55%) at the end of 1994, the last year before the Riegle-Neal Act went into effect. I calculate point-to-polygon distance between

[28] Other authors have gotten around this by assuming that borrowers travel to the branch nearest their house's location.

[29] $Interstate_{B \to C,t} = 1$ for all $t \geq 1995$.

[30] These dates are updated from Amel (1993). Thanks to all four authors for sharing their data.

counties and states using an R-based interface to the Geometry Engine - Open Source project, provided by the **rgeos** package.[31] Specifically, I calculate the distance in miles between each county's population-weighted centroid and the closest border of each state along the surface of the GRS80 ellipsoid, a widely-used approximation of the Earth's shape, using an azimuthal equidistant projection from each county's centroid.

I estimate the reduced-form regression of mortgage sale rates on the two proposed instruments for lender-county pairs between 1990 and 1994. The Douglas Amendment only applies to commercial banks and BHCs, so I exclude subsidiaries of thrift holding companies and non-FDIC-insured HMDA filers (mortgage banks). I also exclude banks owned by foreign banking organizations, since my data do not tell me which state is treated as their home state for interstate banking purposes.

Results are presented in Table 2.5. Column 1 presents a re-estimation of the RDD for loans within $10,000 of the loan limit for the subsample of lenders who are banks or bank holding companies, for loans originated between 1990 and 1994. The controls are the same as in Table 2.3, column 4 and include lender×year fixed effects. The restricted sample is about one-quarter the size of the total sample (104 thousand versus 388 thousand). The estimated coefficient on $\log(HQDist_{bct})$ is 1.38, indicating that a ten percent increase in distance is associated with a 0.138 percentage point increase in the private sale rate. This is about two-thirds the magnitude of the estimate from Table 2.3, column 1, but the 95% confidence intervals for the pair of estimates overlap.

I replace borrower-to-HQ distance with $Interstate_{B\to C,t}$ in column 2 and with the log of $NDNDist_{bct}$ in column 3.[32] The coefficients are large and statistically significant at the 1% level. Column 2 indicates that the private sale rate for borrowers living in a deregulated-neighbor state is 12 percentage points lower than for borrowers living in non-deregulated states (with a 95% confidence interval of -6 to -18 percentage points) for loans originated just above the CLL threshold, but that there is no difference between deregulated and non-deregulated states for loans originated just below the CLL. Column 4 includes both interstate-deregulation-based variables. The coefficient on $\log(NDNDist_{bct})$ is now identified by comparing pairs of borrowers taking out loans from the same lender who both live in non-deregulated states, but at different distances from the nearest deregulated state border. After controlling for the extensive margin of whether or not a borrower lives in a state permitting entry to the BHC, the intensive margin of how far a borrower lives from the nearest deregulated state border appears not to matter.

Column 5 runs the RDD with both HQ distance and the dummy for whether or

[31] Version 0.2-19, Bivand and Rundell (2013).

[32] I bottom-code this distance at 1 mile, so if the borrower lives in a "neighboring" state to the BHC's home state, $log(NDNDist_{bct}) = 0$. All states are neighbors to themselves and are bottom-coded. The minimum distance observed for borrowers living in non-neighboring states is 1.3 miles, so the bottom-coding does not affect distance to non-neighboring states.

not a borrower's home state has deregulated with the BHC's home state. This specification separates lender-borrower distance into two components: whether the borrower could have legally resided near a bank branch, and the residual distance from borrower to bank headquarters after controlling for (an instrument for) borrower-branch distance. Both measures of distance are associated with significantly higher private sale rates for loans made above the CLL, even holding the other constant. These results suggest that frictions in acquiring borrower information and in transmitting this information to higher-ups both affect the loan origination process in the way my model would predict if the lemon effect dominates the quality effect. Secondary-market purchasers are more likely to buy loans that are far away from the bank headquarters and far away from the nearest bank branch. The results also indicate that when the lender can operate a bank branch network in the borrower's home state, it is much less likely to be able to sell the loan. The presence of Fannie and Freddie in the market just below the CLL changes this behavior: private counterparty buyers behave insensitively with regards to both measures of lender-borrower distance.[33]

The results are interesting in light of the move toward interstate deregulation in the 1990s. The Riegle-Neal Act not only allowed a single BHC to own banks chartered in different states; it also allowed the BHC to merge those banks and own a single, nationwide branch network. The estimates in Table 2.5 suggest that increased pervasiveness of local branches negatively impacted lenders' ability to sell loans, consistent with local branches forming relationships with customers and acquiring soft information. On the other hand, the rise of nationwide banking means that branch-to-headquarters distance went up during the 1990s, which tends to reduce loan officers' reliance on soft information but make it easier to sell loans.

2.6.2 Asymmetric Information in the 2000s

In this extension I extend my dataset to include HMDA data between 2001 and 2007. This period spans both the housing boom in the early part of the decade and the decline of 2006-07 which helped precipitate the recession beginning in December 2007. Starting in 2008, the CLL varies by U.S. county and can no longer be treated as exogenous to local economic conditions. I therefore exclude all data from 2008 forward.

To assess whether the nature of strategic bank-purchaser interaction changed, and whether the impact of Fannie and Freddie on players' incentives changed, I run a rolling regression of the RDD around the CLL for every year between 1990 and 2007. Each specification includes bank fixed effects and the same controls for borrower and

[33]In an unreported robustness check, I run the specification with all three variables. The coefficients on HQ distance are similar to column 5 and those on the deregulation variables to column 4. As before, the nearest-neighbor coefficient seems to be driven by whether or not a borrower lives in a deregulated state.

county demographics as reported in Table 2.3, column 4. I plot the coefficients on borrower-to-HQ distance in the conforming market (β^C in equation 2.8) and in the jumbo market (β^J in equation 2.8), along with 95% confidence intervals for each coefficient and for the difference in coefficients (constructed using lender-clustered standard errors). The resulting estimates are depicted in Figure 2.7.

The coefficient on distance in the jumbo-size market is remarkably stable between the years 1990 and 1998, hovering right around 2 for the entire time period. The standard errors are somewhat larger than in Table 2.3 due to the smaller sample size for each year, but even with this reduced power, the coefficient is statistically different from zero at the 5% level for seven of nine years. By contrast, the coefficient on distance in the conforming-size market is negative or close to zero and is insignificant for six of nine years.

Between 1998 and 2000, $\hat{\beta^J}$ dips downwards and becomes about half as large. In the new years added to the sample, 2001-2007, the coefficient appears to be somewhat smaller, between 1 and 2. However, the estimate remains different from zero in all seven new sample years, and the difference $\hat{\beta^J} - \hat{\beta^C}$ is statistically different from zero in all years except for 2002. $\hat{\beta^C}$ is only distinct from zero for one year, 2001.

The graph does not suggest big changes in the regression during the housing boom and bust periods versus the baseline period of the 1990s. There is perhaps some evidence that the coefficient on distance became smaller but remained positve in the jumbo-size market beginnning in 1998. This might be suggestive that the lemon effect became weaker and purchasers were somewhat more worried about the quality effect of whether banks were exerting effort to originate good mortgages, but the timing is not conclusive. However, it is striking that the 2000s look relatively similar to the 1990s. If there was a major change in bank business models in the 2000s towards earning fees by originating and selling a high volume of mortgages, it is not apparent that market participants in the purely-private jumbo segment of the market changed their behavior in response.

2.7 Conclusion

I have shown that the lemon effect dominates the quality effect and mortgage sale rates tend to increase with lender-borrower distance. The regression discontinuity analysis is consistent with Fannie and Freddie reducing lender incentives to acquire borrower private information. Above the CLL, lender-purchaser asymmetric information is very severe and the lemon effect dominates. I find weak evidence of a quality effect in the market segment where the GSEs are present, but this is not robust to the inclusion of lender attributes. The results are robust to allowing the coefficient on distance to vary in a flexible fashion with loan size, so they are probably not driven by omitted borrower attributes which co-vary with loan size. The results are robust to excluding loans taken out within $5,000 of the conforming loan limit,

indicating that my estimates are capturing a more general effect than local borrower or lender sorting around the CLL. Breakpoint tests indicate that it is extremely unlikely we would observe such a large discontinuity in the coefficient on distance by random chance. I find very similar results using measures of lender-borrower distance based on the timing of interstate bank deregulation, which determines whether a BHC was allowed to own a bank branch in any state-year.

Put succinctly, this is evidence that the GSEs ameliorate the lemons problem by reducing bank information production. This has important implications both for our interpretation of the financial crisis and for reforming the mortgage market in its wake. In terms of the financial crisis, my results indicate that private markets were wholly capable of developing a mechanism to cope with the lemons problem. Fannie and Freddie distorted this mechanism by encouraging a model of informationally-insensitive, fee-based originations in which the lender exerted minimal effort and retained no stake in the loans. It is hard not to draw parallels between this description of the conforming market in the 1990s and the rise of subprime securitization in the 2000s. If the current policy debate on the future of Fannie and Freddie were to result in a purely private market, we might expect it to look much like the jumbo market in the 1990s. In addition to mortgage credit supply decreasing and prices increasing, informational asymmetries between originators and purchasers would likely become more severe. However, jumbo mortgage markets were not local and isolated. Mortgage credit continued to flow to communities far away from the lender's headquarters or branch network, and private counterparties were willing to finance these loans by purchasing them on secondary markets. This offers some hope for a future where Fannie and Freddie are eliminated or their role is greatly reduced.

Figure 2.1

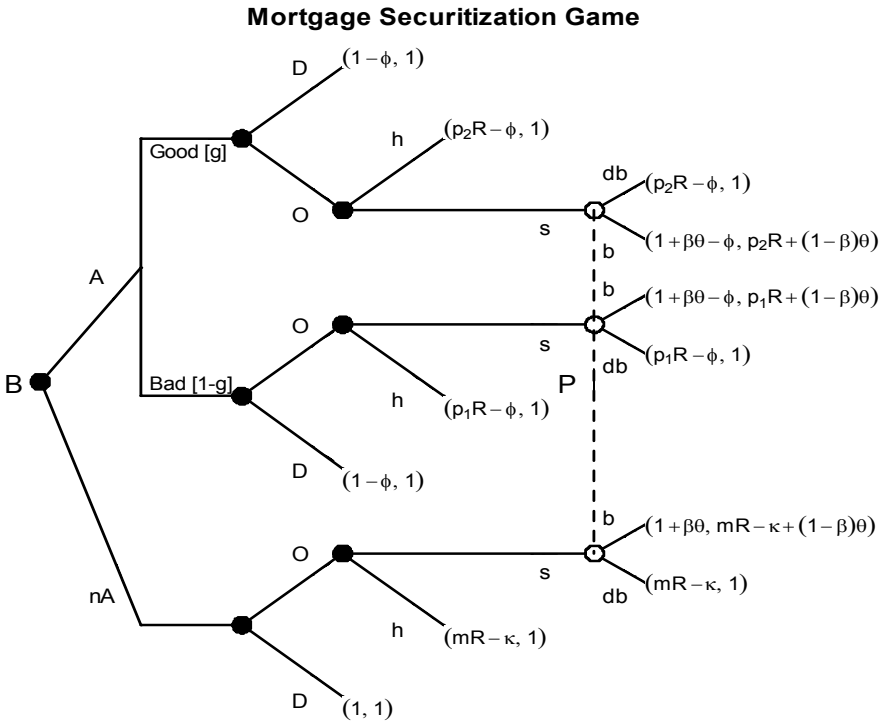
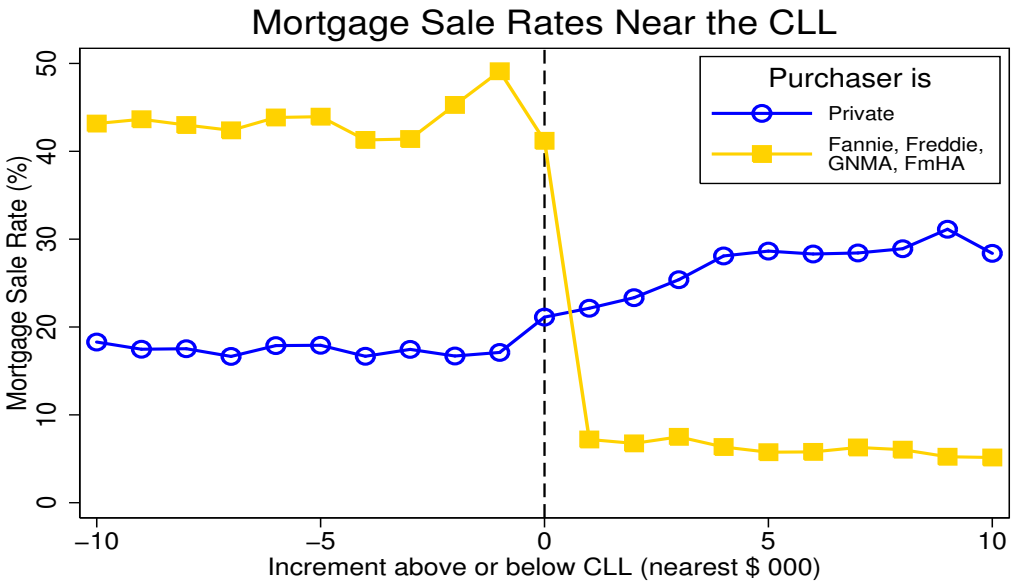

Mortgage Securitization Game
B
A
nA
Good [g]
Bad [1-g]
D
O
h
s
b
db
P
$(1-\phi, 1)$
$(p_2R-\phi, 1)$
$(1+\beta\theta-\phi, p_2R+(1-\beta)\theta)$
$(1+\beta\theta-\phi, p_1R+(1-\beta)\theta)$
$(p_1R-\phi, 1)$
$(1+\beta\theta, mR-\kappa+(1-\beta)\theta)$
$(mR-\kappa, 1)$
(1, 1)

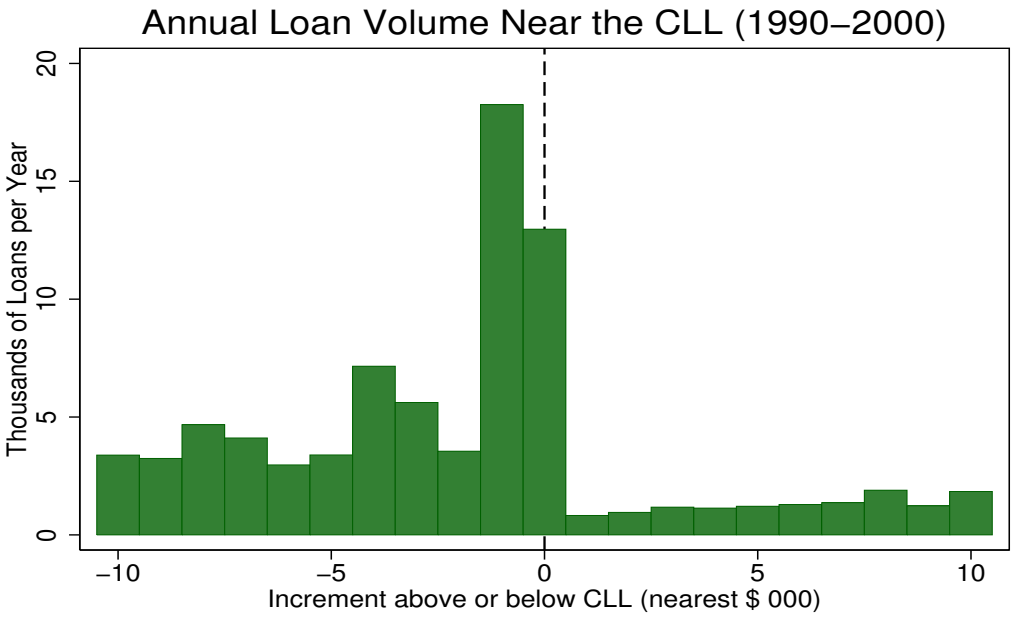
Figure 2.2



Source: HMDA (1990–2000) / author's calculations.

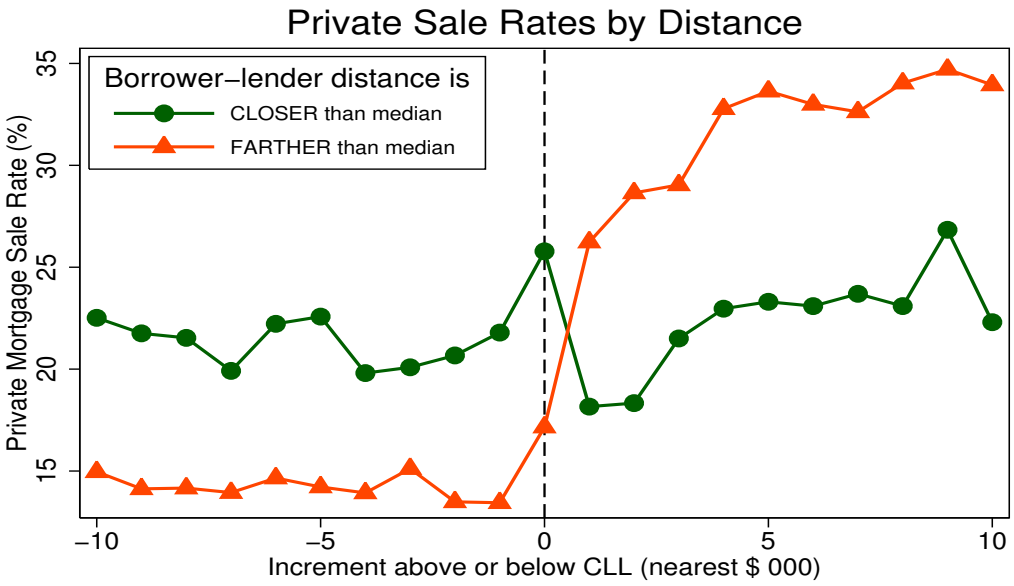
Average across lender–county–loan size cells, 1990–2000.

Figure 2.3



Source: HMDA / author's calculations.

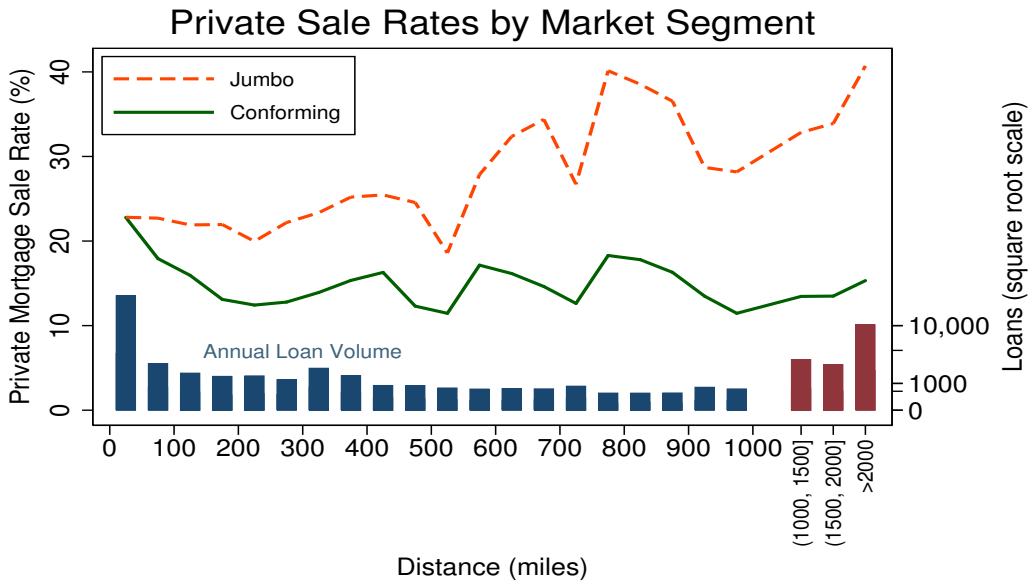
Figure 2.4



Source: HMDA (1990–2000) / author's calculations.

Average across lender–county–loan size cells, 1990–2000.

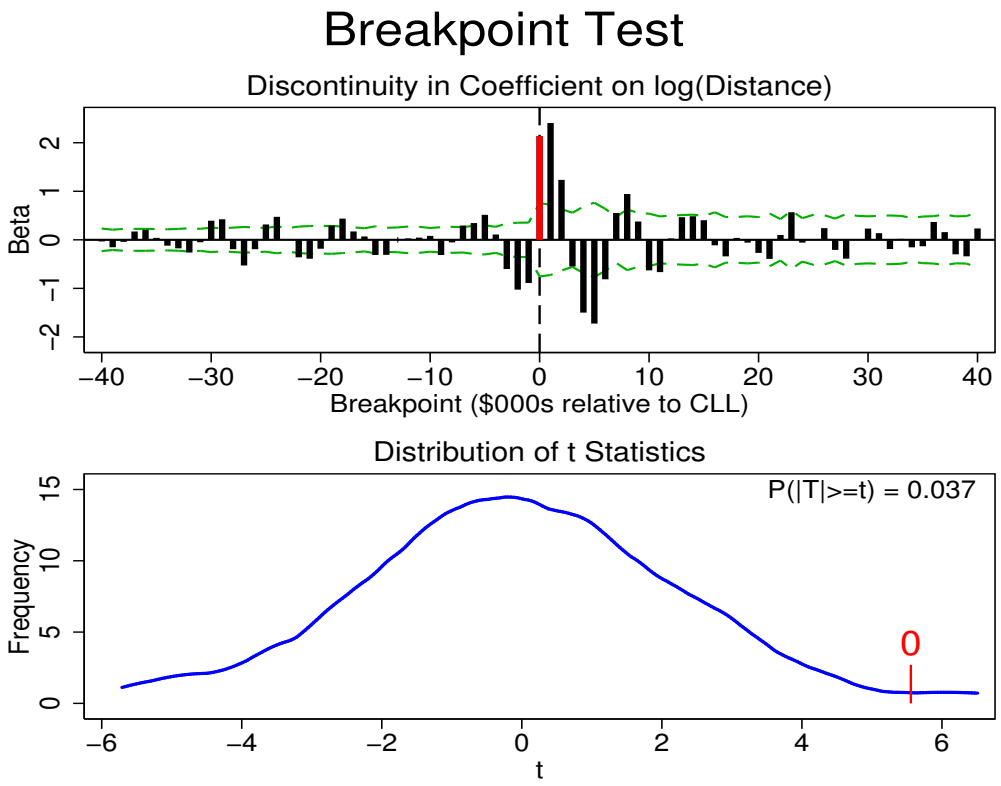
Figure 2.5



Source: HMDA (1990–2000) / author's calculations.

Average across lender–county–loan size cells, 1990–2000. Zero cell excluded.

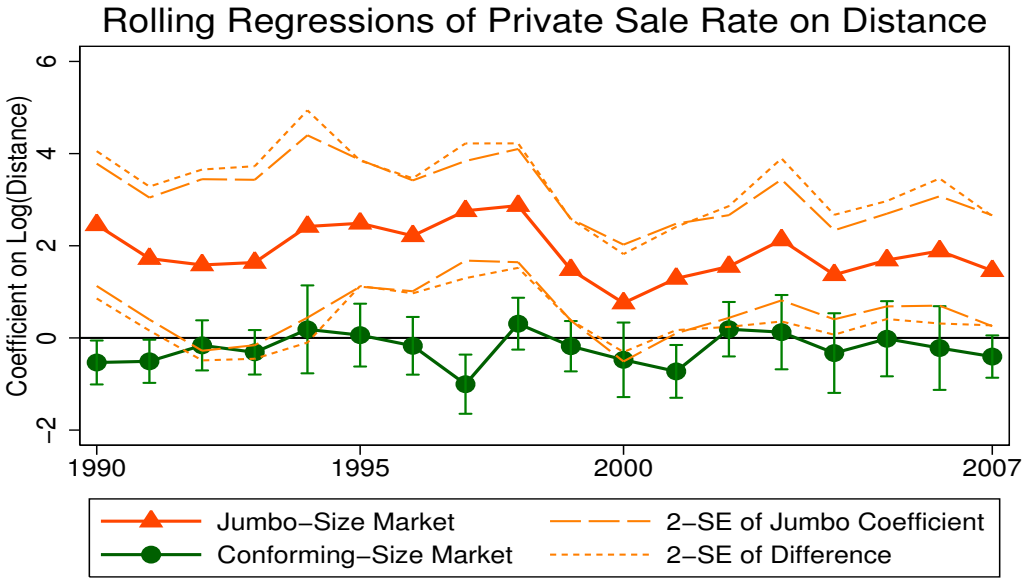
Figure 2.6



Source: HMDA (1990–2000) / author's calculations.

Rolling RDDs in [–$10K, +$10K] window around breakpoint, controlling for P(5) of loansize × log(distance).

Figure 2.7



Source: HMDA / author's calculations.

RDD within $10K of CLL, with borrower & county controls + lender FX.

Table 2.1: Greater Distance Predicts a Higher Mortgage Sale Rate

OLS panel regression on Lender × continental U.S. Counties in MSAs, 1990-2000. [1]

Dependent variable is percent of mortgages the lender sells:	*Overall*		*If borrower is in Bottom State Income Quintile [2]*		*Bottom - Top State Income Quintile [2]*	
	(1)	(2)	(3)	(4)	(5)	(6)
Log(Distance from Borrower to Lender HQ) [3]	2.127***	-0.13	1.882***	1.779***	0.618***	1.087***
	(0.44)	(0.15)	(0.70)	(0.29)	(0.21)	(0.20)
Lender Competition Index by # of Applications (0-100)	-0.0719	0.0129	-0.236	-0.0103	-0.0627**	-0.0413*
	(0.060)	(0.011)	(0.16)	(0.019)	(0.031)	(0.022)
Log(County Population Density in Census 2000)	0.0907	-0.0472	2.135***	0.712***	0.442**	0.275
	(0.22)	(0.090)	(0.52)	(0.23)	(0.19)	(0.23)
Log(County Median Household Income in 2000 in $)	13.08***	2.443***	18.41***	8.322***	2.418	5.210**
	(1.97)	(0.92)	(4.80)	(2.03)	(1.92)	(2.51)
Growth Rate of Median HH Income, 1990-2000 (%)	0.0295	-0.0274***	0.00481	-0.0375	-0.00121	0.0135
	(0.024)	(0.009)	(0.047)	(0.023)	(0.020)	(0.020)
Gini Income Concentration Index in County (0-100)	0.597***	0.0278	0.869***	0.0832	0.0312	0.0285
	(0.069)	(0.033)	(0.130)	(0.076)	(0.054)	(0.068)
State House Price Growth Rate from FHFA (%)	0.0889	-0.0115	0.341**	0.139	0.187**	0.240**
	(0.090)	(0.055)	(0.17)	(0.12)	(0.077)	(0.10)
Percent Jumbo Loans in County	-0.321***	-0.184***	-0.199	-0.248***	0.130***	0.100**
	(0.046)	(0.020)	(0.120)	(0.050)	(0.042)	(0.045)
Borrowers' Loan-to-Income Ratio	2.356***	0.714***	0.715	0.560***	0.644***	0.679***
	(0.500)	(0.190)	(0.500)	(0.170)	(0.230)	(0.240)
Borrowers' Income Percentile Rank in County (0-100)	0.00689	-0.00573	-0.12	0.123***	0.102*	0.205***
	(0.050)	(0.014)	(0.160)	(0.037)	(0.055)	(0.060)
Number of States in which Lender Operates	-0.260***		-0.253**		-0.00721	
	(0.085)		(0.12)		(0.034)	
Log(Total # of Mortgages Lender Originates)	5.782***		6.047***		0.179	
	(0.58)		(1.03)		(0.27)	
Year FX	YES	YES	YES	YES	YES	YES
Lender FX		YES		YES		YES
Lender×Year FX		YES		YES		YES
Observations	893,601	893,601	166,716	166,716	130,595	130,595
R-squared	0.12	0.73	0.13	0.69	0.01	0.24

Standard errors clustered by lender in parentheses
*** p<0.01, ** p<0.05, * p<0.1

Notes.

[1] HMDA reporters held by the same bank holding company are aggregated to form a single "lender."

[2] State income quintile cutoffs are estimated from Small Area Income and Poverty Estimates, assuming a gamma distribution. See main text for details.

[3] City where lender is headquartered to population-weighted county centroid where borrower's property is located. Great circle distance is calculated using the spherical law of cosines.

Table 2.2: Do Agents Sort on the CLL?

Difference in mean values of attributes (above CLL - below CLL) after year FX, with S.E.s.

Other Controls:	$\mathcal{P}(5)$ of Loan Size	
Dependent Variable (Below)	(1)	(2)
A. County Characteristics [1]		
Median HH Income in 2000	0.26	0.179
($ 000s)	(0.74)	(0.33)
Gini Income Concentration Index	1.150***	0.533***
(0-100)	(0.36)	(0.15)
Population Density in 2000	1588**	740.6**
(people / sq. mile)	(624)	(290)
B. Borrower Characteristics [1]		
Income Percentile Rank in County	-4.020***	0.318**
(0-100)	(0.38)	(0.14)
Loan-to-Income Ratio	0.230***	0.0572***
	(0.036)	(0.011)
C. Lender Characteristics [2]		
Total # of Mortgages Lender	0.344	0.308
Originates	(1.57)	(1.14)
Number of States in which	-2.115***	-0.35
Lender Operates	(0.80)	(0.52)

*** p<0.01, ** p<0.05, * p<0.1

Notes.

Each observation is a Lender x County x Loan Size Buckets in $000s. Sample is [CLL -$10K, CLL + $10K], excluding zero cell. T=1990-2000.

[1] Standard errors clustered by county are in parentheses.

[2] Standard errors clustered by lender are in parentheses.

Table 2.3: Regression Discontinuity Analysis Using the CLL

OLS panel regression on Lender × continental U.S. Counties in MSAs × Loan Size Buckets in $000s, [CLL-$10K, CLL + $10K], zero cell omitted. T = 1990-2000. [1]

Dependent variable:	*Percent of Mortgages Sold to Private Counterparties*			
	(1)	(2)	(3)	(4)
Log(Distance from Borrower to Lender HQ) [2]	-2.075***	-2.260***	-0.164	-0.186
	(0.31)	(0.29)	(0.26)	(0.18)
1{Jumbo} × Log(Distance)	3.060***	3.006***	2.917***	2.118***
	(0.50)	(0.51)	(0.51)	(0.48)
1{Jumbo Mortgage}	-7.645***	-7.879***	-7.006***	0.199
	(1.33)	(1.35)	(1.38)	(1.26)
Lender Competition Index by # of Originations (0-100)		0.211***	0.130**	0.0821***
		(0.059)	(0.051)	(0.024)
Log(County Population Density in Census 2000)		-0.803***	-0.438*	-0.146
		(0.26)	(0.230)	(0.110)
Log(County Median Household Income in 2000 in $)		12.55***	12.40***	4.758***
		(1.86)	(1.820)	(0.690)
Growth Rate of Median HH Income, 1990-2000 (%)		-0.000237	0.0208	-0.0258*
		(0.037)	(0.035)	(0.014)
Gini Income Concentration Index in County (0-100)		0.995***	0.968***	0.190***
		(0.120)	(0.120)	(0.041)
State House Price Growth Rate from FHFA (%)		0.369***	0.371***	0.0669
		(0.110)	(0.099)	(0.061)
Loan-to-Income Ratio		0.226	0.191	0.066
		(0.190)	(0.150)	(0.064)
Income Percentile Rank in County (0-100)		-0.0117	-0.0172	0.0382***
		(0.022)	(0.020)	(0.009)
Number of States in which Lender Operates			-0.0249	
			(0.067)	
Log(Total # of Mortgages Lender Originates)			-2.956***	
			(0.560)	
Year FX	YES	YES	YES	YES
Lender FX				YES
Lender×Year FX				YES
F test, Dist coeff above CLL = 0	2.863	1.647	35.69	24.41
Prob > F	0.0907	0.199	0.000	0.000

Observations	396,645	387,624	387,624	387,624
R-squared	0.03	0.04	0.06	0.56

Standard errors clustered by lender in parentheses
*** p<0.01, ** p<0.05, * p<0.1

Notes.

[1] HMDA reporters held by the same BHC are aggregated to form a single "lender.

[2] City where lender is headquartered to population-weighted county centroid where borrower's property is located. Great circle distance is calculated using the spherical law of cosines.

Table 2.4: RDD Robustness Checks

OLS panel regression on Lender × continental U.S. Counties in MSAs × Loan Size Buckets in $000s, [CLL-$10K, CLL + $10K], zero cell omitted. T = 1990-2000. [1]

Dependent variable:	*Percent of Mortgages Sold to Private Counterparties*			
	(1)	(2)	(3)	(4)
Log(Distance from Borrower to Lender HQ) [2]	-0.184	[5]	-0.234	-0.313
	(0.18)		(0.20)	(0.23)
1{Jumbo} × Log(Distance)	2.101***	2.137***	2.089***	2.082***
	(0.48)	(0.38)	(0.49)	(0.54)
1{Jumbo Mortgage}	-3.050*	-3.278**	0.642	1.927
	(1.56)	(1.29)	(1.29)	(1.47)
Sample excludes?	zero cell	zero cell	CLL ± $1K	CLL ± $5K
$\mathcal{P}(5)$ of Loan Size	YES	YES		
$\mathcal{P}(5)$ of Loan Size × Log(Distance)		YES		
County Controls [3]	YES	YES	YES	YES
Borrower Controls [4]	YES	YES	YES	YES
Year FX	YES	YES	YES	YES
Lender FX	YES	YES	YES	YES
Lender×Year FX	YES	YES	YES	YES
Observations	387,624	387,624	318,169	171,122
R-squared	0.56	0.56	0.56	0.56

Standard errors clustered by lender in parentheses
*** p<0.01, ** p<0.05, * p<0.1

Notes.

[1] HMDA reporters held by the same BHC are aggregated to form a single "lender."

[2] City where lender is headquartered to population-weighted county centroid where borrower's property is located. Great circle distance is calculated using the spherical law of cosines.

[3] County controls are: lender competition index, log(population density in 2000), log(median HH income in 2000) and 1990-2000 growth rate, Gini income concentration index, and state house price growth rate.

[4] Borrower controls are: loan-to-income ratio and borrower's income percentile ranking in county.

[5] Coefficient represents the slope on log(distance) for a loan of size 0, so has no economic interpretation. Slope = -0.353, S.E. = 0.18.

Table 2.5: RDD with an Alternate Measure of Distance

OLS panel regression on Commercial Bank or BHC × continental U.S. County in MSA × Loan Size Bucket in $000s, [CLL-$10K, CLL + $10K], zero cell omitted. T = 1990-1994. [1]

Dependent variable:	*Percent of Mortgages Sold to Private Counterparties*				
	(1)	(2)	(3)	(4)	(5)
Log(Distance from Borrower to Lender HQ) [2]	0.0525				0.131
	(0.17)				(0.17)
1{Jumbo} × Log(HQ Distance)	1.382***				0.832**
	(0.40)				(0.38)
1{Borrower Lives in a Deregulated State} [3]		1.153		0.814	1.228
		(0.80)		(1.59)	(0.82)
1{Jumbo} × 1{Borrower Lives in a Deregulated State}		-12.45***		-26.00***	-10.08***
		(3.07)		(9.16)	(3.01)
Log(Distance from Borrower to Nearest Deregulated State)			-0.244	-0.0812	
			(0.17)	(0.38)	
1{Jumbo} × Log(NDS Distance)			2.231***	-2.779	
			(0.61)	(1.71)	
1{Jumbo Mortgage}	1.82	19.08***	6.793***	32.63***	13.30***
	(1.24)	(3.40)	(1.13)	(9.45)	(3.36)
County Controls [4]	YES	YES	YES	YES	YES
Borrower Controls [5]	YES	YES	YES	YES	YES
Year FX	YES	YES	YES	YES	YES
Lender FX	YES	YES	YES	YES	YES
Lender×Year FX	YES	YES	YES	YES	YES
Observations	103,966	103,966	103,966	103,966	103,966
R-squared	0.48	0.48	0.48	0.48	0.48

Standard errors clustered by lender in parentheses
*** p<0.01, ** p<0.05, * p<0.1

Notes.

[1] HMDA reporters held by the same BHC are aggregated to form a single "lender." Only commercial bank or BHC high-holders are included in sample.

[2] City where lender is headquartered to population-weighted county centroid where borrower's property is located. Great circle distance is calculated using the spherical law of cosines.

[3] "Deregulated states" are states which allow interstate banking entry for BHCs from lender's home state, including lender's home state. Distance is measured from borrower's population-weighted county centroid to state borders. Distance coded as "1 mile" if borrower's state permits entry from lender's state.

[4] County controls are: lender competition index, log(population density in 2000), log(median HH income in 2000) and 1990-2000 growth rate, Gini income concentration index, and state house price growth rate.

[5] Borrower controls are: loan-to-income ratio and borrower's income percentile ranking in county.

Chapter 3

The Welfare Consequences of Experienced Inflation on Residential Mortgage Choice

3.1 Introduction

Whether to buy a home and how to finance the purchase is one of the biggest financial decisions most households will face in their lifetimes. The dominant contract type in the United States is the 30-year, level-payment, self-amortizing, fixed rate mortgage (hereafter referred to as "FRM"). This contract type's popularity was encouraged by the Congress's establishment of Fannie Mae in 1938 and Freddie Mac in 1970 with the mission of purchasing long-term fixed rate mortgages from banks which might otherwise face duration risk from holding these assets. Following the onset of the S&L crisis, the Garn-St. Germain Depository Institutions Act of 1982 allowed banks to originate adjustable-rate mortgages ("ARMs"). A typical ARM contract still self-amortizes over a long-term period such as 30 years, but the interest rate resets periodically according to a prespecified margin over an index, typically a one-year Treasury or a district cost-of-funds index, so the monthly payments may vary from year to year.[1] Despite their greater liquidity on secondary mortgage markets, FRMs are priced at a premium over ARMs, in part because they provide insurance against nominal interest rate fluctuations. Freddie Mac's Primary Mortgage Market Survey reports that FRMs carried an average premium of 170 basis points over equivalent credit risk and term ARMs between 1984 and 2013.[2] Figure 3.1 shows the correlation between the FRM-ARM spread and owner-occupied residential mortgage choice based on data of outstanding residential mortgages in 1991 and 2001 collected by the Census Bureau. Fixed rate mortgages have rarely commanded less than 80% market share.

A recent literature in psychology and economics suggests that individuals overweight recent experiences relative to the optimal Bayesian scheme (see Malmendier and Nagel 2013 and the references therein). For example, young borrowers coming of age during the 1970s have recently experienced a period of high inflation, and they do not have personal memory of earlier periods of lower inflation. If this high experienced inflation translates into a forecast of high future inflation, then these borrowers might demand greater insurance against volatility in nominal interest rates. The testable prediction is that mortgagors who belong to younger cohorts in the 1980s should "ignore" price signals and be more likely to choose fixed-rate mortgages, while younger mortgagors in the 1990s who came of age after the Volcker Fed tamed inflation should behave more like older cohorts who came of age prior to the Great Inflation.

[1] More exotic mortgage types became popular in the housing boom period of the 2000s – including "hybrid ARMs" whose interest rate are initially fixed but then become variable, and "interest-only" mortgages in which no principal is paid in early periods to keep initial payments low – but these products are outside the scope of this paper.

[2] The annual average spread fluctuated between 67 and 302 basis points over this time period (S.D. = 67 basis points).

3.2 Data and Estimation

I use annual CPI-U data to calculate experienced inflation $\pi^e_{s,t}$ in year t for individuals belonging to the cohort born in year s using linearly increasing weights:

$$\pi^e_{s,t} \equiv \sum_{k=s}^{t} \frac{k-s}{\sum_{j=s}^{t}(j-s)} \cdot \pi_k \tag{3.1}$$

This formula places zero weight on actual inflation in years prior to an individual's birth and linearly increasing weights on more recent inflation experiences. Malmendier and Nagel (2011, 2013) find that individuals' self-reported inflation expectations in the Michigan Survey of Consumer Confidence follow a pattern very close to Equation 3.1. Figure 3.2 plots experienced inflation and mortgage product choice for individuals under 35 and over 45 for 1985-1991 and 1995-2001. In the late 1980s, younger cohorts had experienced higher rates of inflation and were more likely to choose fixed-rate products than older cohorts. In the late 1990s experienced inflation across younger and older cohorts converged; at the same time, mortgage product choice also converged.

The Census Bureau formerly conducted a Residential Finance Survey the year after each Census year.[3] The RFS consisted of two cross-referenced surveys, one to households and one to their mortgage lenders. The household arm of the survey provides household demographic and income data, while the lender arm provides terms on any outstanding loans secured by the property. The sample is drawn from the Census roster of households, so there is a tendency to miss households that have recently moved. The sample scheme oversamples multi-unit properties, particularly rental properties with 5+ units, but it is otherwise designed to be representative of the stock of outstanding mortgages in the preceding Census year. I obtain microdata on the mortgages linked to owner-occupied 1-4 unit properties from the 1991 and 2001 waves of the RFS. Since the sample is of outstanding mortgages, I am missing mortgages that were refinanced, prepaid, or defaulted upon prior to the survey year. To minimize these issues and approximate a flow dataset of mortgage choice situations, I restrict the sample to mortgages which were taken out no more than six years prior to the survey year (1985-1991 and 1995-2001, respectively).[4] For households with multiple members, I use the age of whoever self-identifies as the primary owner. Total household income in the survey year is imputed back to the origination year by peak-to-peak log growth rate in U.S. nominal median household income over 1980-2001 from CPS Historical Table H-6 (approximately 4.14% annually).

The RFS consistently defines three types of mortgage products across both survey waves: the aforementioned FRM and ARM alternatives, and balloon mortgages. This third alternative features level payments over the life of the loan which are

[3]The RFS was unfortunately discontinued prior to the 2010 Census.

[4]In the 1991 survey, origination years are only reported in intervals: 1985-86, 1987-88, and 1989-91.

not fully amortizing, so a large lump or "balloon" payment of the remaining principal is due at maturity, usually 7-10 years. Balloon mortgages are designed to attract borrowers who would not otherwise qualify for a fully-amortizing product. Balloon mortgages offer lower monthly payments and the borrower may be able to refinance upon maturity if his situation has improved, but they carry greater risk as the borrower will have to default if he cannot refinance and cannot afford the balloon payment (MacDonald and Holloway 1996). Borrower attributes are summarized by mortgage product choice in Table 3.1. Borrowers choosing ARMs tend to be higher income and are less likely to be first-time homeowners. Experienced inflation for both groups is higher than contemporaneous (actual) inflation; for borrowers choosing an FRM, $\pi^e - \pi = 4.77 - 3.38 = 1.39$ percentage points, while for borrowers choosing an ARM, $\pi^e - \pi = 4.81 - 3.47 = 1.34$ percentage points. Borrowers choosing an FRM thus have a larger over-estimate of inflation, relative to its current level, than do borrowers choosing an ARM.

I use this data to estimate a McFadden-style model of residential mortgage choice. The model specifies that the household in choice situation n derives utility $U_{ni} = x'_{ni}\beta + \varepsilon_{ni}$ from alternative $i \in \{FRM,\ ARM,\ Balloon\}$. Alternative i is chosen if $U_{ni} > U_{nj}$ for all $j \neq i$. By assuming that attribute characteristics which are not observed by the econometrician, ε_{ni}, follow a Type I extreme value distribution, Marley (cited by Luce and Suppes 1965) and McFadden (1974) showed that choice probabilities may be described by a logit formula whose likelihood function is globally concave, so the appropriately-standardized utility parameters may be easily estimated by maximum likelihood.

Theoretically, the mortgage payment structure preferred by a household will depend on its age and mobility, current and expected future income, risk aversion, and its beliefs about future short-term interest rates (see, among others, Stanton and Wallace 1998, Campbell and Cocco 2003, Chambers et al. 2009, and Koijen et al. 2009). Writing this down in indirect utility terms:

$$U_{ni} = \alpha_{it} + \beta_R Rate_{ni} + \beta_{\pi,i}\pi^e_n + \beta_{Inc,i} Income_n + f_i(Age_n) + \varepsilon_{ni} \tag{3.2}$$

The observable components of the utility an individual derives from alternative i depend on the interest rate of that alternative, the borrower's income, the borrower's age, and the borrower's experienced inflation. Alternative-specific year fixed effects α_{it} control for the overall (un-)desirability of a given alternative in a given year, so capture the rational-expectations forecast about the economic environment which should be common to all households at any point in time. In the presence of year fixed effects, a borrower's recent inflation experiences should not matter, unless there is a correspondence between those experiences and borrower beliefs which differ from the rational-expectations forecast. Specifically, I predict that $\beta_R < 0$ and $\beta_{\pi,FRM} > 0$. (Only differences in utility affect choice probabilities, so I normalize $\beta_{\cdot,ARM} \equiv 0$ for all sociodemographic characteristics, including experienced inflation).

There are two wrinkles to estimating this choice problem. First, the interest rates

of the non-chosen alternatives are not observed. Since the pattern of missing data is "missing at random" (following Rubin's 1976 nomenclature), imputation-based methods are available. To one degree or another, all these methods estimate the correlations between observed borrower characteristics and interest rates using the subsample of borrowers who chose each alternative, then use these characteristics and estimated parameters to sample from the distribution of interest rates for the non-chosen alternatives. Essentially, one would estimate the parameters γ_i of the following equation:

$$Rate_{ni}^{chosen} = z_n'\gamma_i + u_{ni}^{chosen} \tag{3.3}$$

This introduces the second wrinkle. Unobserved factors that determine an individual's choice set are unlikely to have conditional mean equal to zero; rather, it seems likely that $\mathbb{E}[u_{ni}^{chosen}|z_n] < 0$. Borrower selection or lender screening based on unobserved variables will pose an external validity problem when the estimated parameters are used to impute the interest rates of the non-chosen alternatives. Lee (1978) confronted a similar problem with regards to estimating the wages of union versus non-union jobs, and Brueckner and Follain (1988) first applied Lee's methodology to a mortgage choice setting. The key insight is that just as the errors of the chosen alternative are likely to be negative, the errors of the non-chosen alternatives are likely to be positive. Plugging the exogenous variables z_n from equation 3.3 into the utility equation 3.2, one obtains that alternative i is chosen if $\varepsilon_{nj} - \varepsilon_{ni} > \alpha_{it} - \alpha_{jt} + \beta_R z_n'(\gamma_i - \gamma_j) + \beta_R(u_{ni} - u_{nj}) + ...$ for all $j \neq i$. It is plausible to assume that the errors u_{ni} and u_{nj} cancel out, so the reduced form choice model may be estimated consistently.

Freddie Mac's Primary Mortgage Market Survey provides weekly data on average FRM and ARM interest rates from a representative nationwide sample of mortgage originators. The representative products are first-lien, prime, conventional, conforming mortgages with an LTV of 80% and a 30-year term. I re-weight from the five Freddie Mac regions to the four Census regions using 1990 Census housing units by state and take annual averages.[5] The reduced form of utility that individual n derives from alternative i, residing in Census region r in year t, is thus

$$U_{ni} = \alpha_{it} + \tilde{\beta}_R PMMSRate_{r,t,i} + \beta_{\pi,i}\pi_n^e + \beta_{Inc,i}Income_n + f_i(Age_n) + \tilde{\varepsilon}_{ni} \tag{3.4}$$

The estimation sample is borrowers aged 25-74 in the year of origination (restricted to 1985-91 and 1995-2001, respectively) for whom all covariates are available. Table 3.2 presents estimates of the multinomial logit model. Each coefficient represents that attribute's or sociodemographic characteristic's contribution to the utility of

[5]The RFS reports the home state of borrowers residing in a few large states. In these cases I simply use the corresponding Freddie Mac region interest rate.

that alternative. So, for example, $\hat{\tilde{\beta}}_R = -0.424$ in column 1, indicating that individuals derive less utility from and are less likely to choose more expensive alternatives. All columns include alternative-specific year fixed effects and control for a quadratic function of the primary owner's age. Column 1 estimates a single price coefficient on both the FRM and the ARM initial rate indices (so only the spread matters), while columns 2-4 allows the two coefficients to differ. Column 3 normalizes $\beta_{\pi,Balloon} = \beta_{\pi,ARM}$, while column 4 controls for characteristics of the mortgage (seniority, whether it is a refinancing of a previous mortgage, conventional dummy, and points paid).

The results indicate indicate that individuals who have higher levels of π^e as of the year of the choice situation derive greater utility from the FRM alternative, relative to the omitted ARM alternative. Experienced inflation reduces the utility of a balloon mortgage relative to an ARM, but this effect is imprecisely estimated and not significant at standard levels. A useful normalization is to calculate the compensating interest rate differential an individual is willing to pay for one additional percentage point of experienced inflation. The estimates in column 1 indicate that individuals are willing to pay $0.211/0.424 = 0.498$ percentage points in the FRM - ARM spread due to an additional percentage point of π^e. Column 2 indicates that individuals are more sensitive to the fixed rate component of the spread: individuals are willing to pay $0.208/3.57 = 0.058$ percentage points more in the FRM rate due to an additional percentage point of π^e. Since all specifications include origination year fixed effects, these effects are above and beyond the full-information inflation expectation for a given year. Fully rational individuals should place a weight of zero on their personally experienced inflation. Instead, we observe that individuals who have experienced relatively higher levels of inflation derive greater utility from the fixed-rate, inflation-insured alternative.

Figure 3.3 plots the counterfactual FRM share we would observe if individuals ignored π^e. I estimate counterfactual probabilities that an individual would pick each alternative using the coefficients from Table 3.2, column 3, except that I force the coefficient $\beta_{\pi,FRM} = 0$, and aggregate these probabilities to calculate hypothetical product shares for each origination year. In 1985-86, I predict that the FRM share would have been 29 percentage points lower (53% rather than 82%), while the ARM share would have been 22 percentage points higher (42% rather than 16%). The effect of experienced inflation diminishes over time: by 2001, the counterfactual FRM share is 19 percentage points lower than the actual share (62% rather than 83%).

3.3 Welfare

While the effect of experienced inflation on mortgage product shares appears to be economically large, it is not obvious that this is a costly mistake. Figure 3.4 plots the path of the PMMS FRM index, ARM initial index, and a 1-year constant

maturity Treasury plus a margin of 2.75 percentage points during the two RFS sample periods. Although individuals might initially save by choosing an ARM, resets into higher rates could eliminate these savings, particularly in years such as 1988, 1989, and 1996-2000.

To simulate the welfare consequences of mortgage choice on monthly payments, I assume that all mortgages are originated on January 1, carry a 30-year term, are self-amortizing, and are paid on time; that individuals choosing an FRM will receive their regional PMMS rate; that individuals choosing an ARM will receive their regional ARM rate for the first year; and that resets will occur every year based on the average value of the 1-year constant maturity Treasury for that year plus a margin of 2.75 percentage points. The most obvious limitation to this exercise is that I do not adjust interest rates for each individual's risk characteristics. Depending on how risk is priced into initial rates versus margins, this could have an ambiguous effect on the results. On the other hand, most ARMs carry caps on both annual resets and lifetime resets, limiting the amount of interest rate risk borne by the consumer. This will tend to underestimate the potential savings from choosing an ARM over an FRM.

I calculate each individual n's expected annual payment as

$$PMT_n = \frac{p_{n,FRM} \cdot PMT_{n,FRM} + p_{n,ARM} \cdot PMT_{n,ARM}}{p_{n,FRM} + p_{n,ARM}}$$

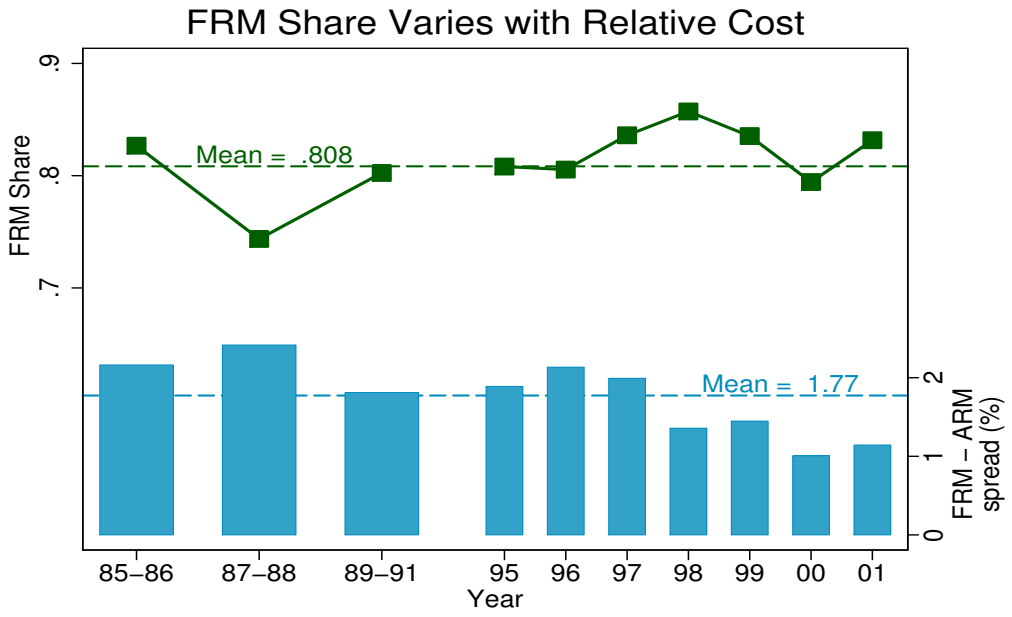
so the balloon alternative is ignored. Actual and counterfactual probabilities are calculated using the coefficients estimated in Table 3.2, column 3. Each individual's actual mortgage principal is used (in 2000 dollars). Figure 3.5 plots the extra annual interest paid under the true probabilities minus the counterfactual probabilities. Potential savings are largest in the initial year, while potential savings are negative in years such as 1989 when short-term interest rates are high. Averaging over all choice situations and years, individuals pay $220 more in mortgage interest per year using the choice probabilities which put a positive weight on experienced inflation.

A second useful exercise is to accumulate the foregone interest rate savings over time. All of the mortgage originations in my sample were still outstanding as of 1991 and 2001, so the accumulated excess interest indicates how costly it has been for each set of borrowers to hold on their mortgage and not refinance. On average, borrowers in the 1991 RFS had cumulatively paid $1,238 in extra interest due to experienced inflation (as of year-end 1991), and borrowers in the 2001 RFS had cumulatively paid $290 extra (as of year-end 2001). Figure 3.6 shows these results graphically. Only borrowers taking out mortgages in 1998, the year with the lowest average FRM rate in sample, do better with their true choice probabilities than by "ignoring" their experienced inflation.

3.4 Conclusion

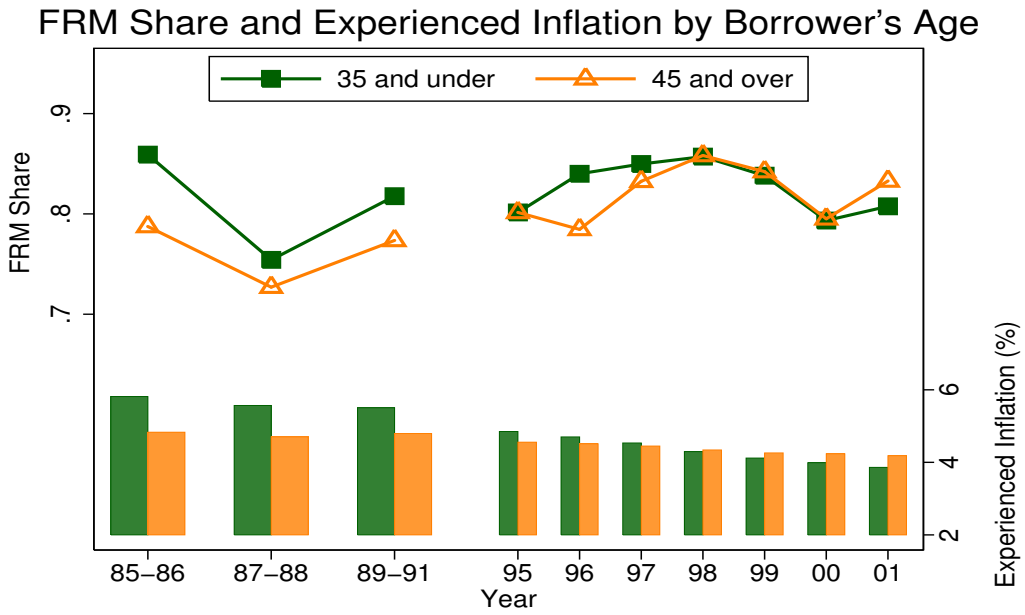
The observation that personal experiences matter continues to bear useful fruit in economics. This chapter has shown that experienced inflation has a significant impact on residential mortgage product choice: individuals experiencing higher levels of inflation are *ex ante* willing to pay more for a fixed-rate product offering insurance against nominal interest rate fluctuations. Moreover, these mistakes are *ex post* costly: a simple simulation exercise suggests that consumers taking out and holding mortgages between 1985-91 and 1995-2001 overpaid by an average of $220 in mortgage interest per year by putting positive weight on their personal inflation experiences. The main source of variation in this chapter is birth year, but additional sources of variation might be found by looking at immigrant groups from countries with different inflation experiences. Additionally, it would be interesting to know whether lenders are aware of and exploit this apparent bias in consumer behavior.

Figure 3.1



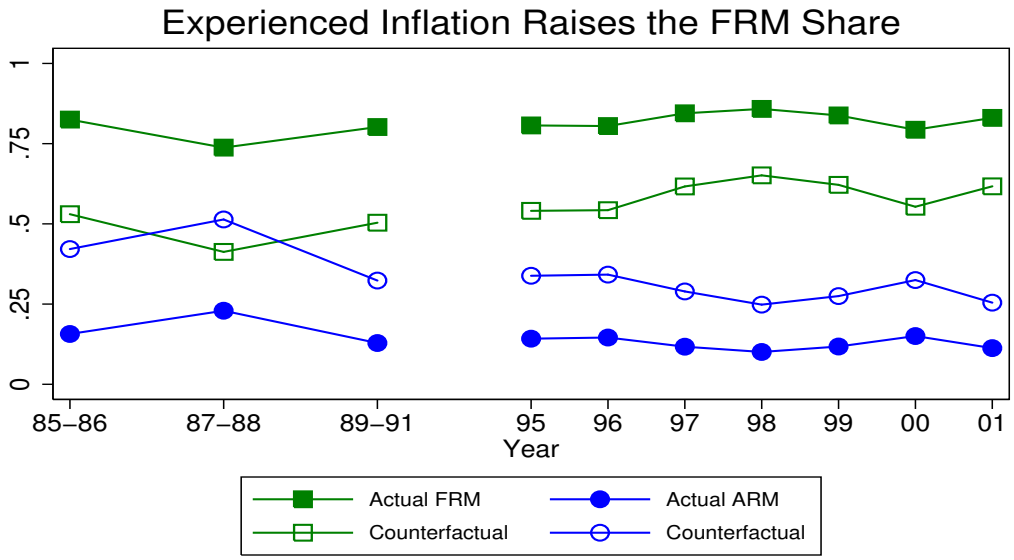
Sources: 1991 & 2001 RFS, Freddie Mac PMMS / author's calculations.

Figure 3.2



Sources: 1991 & 2001 RFS, BLS CPI / author's calculations.

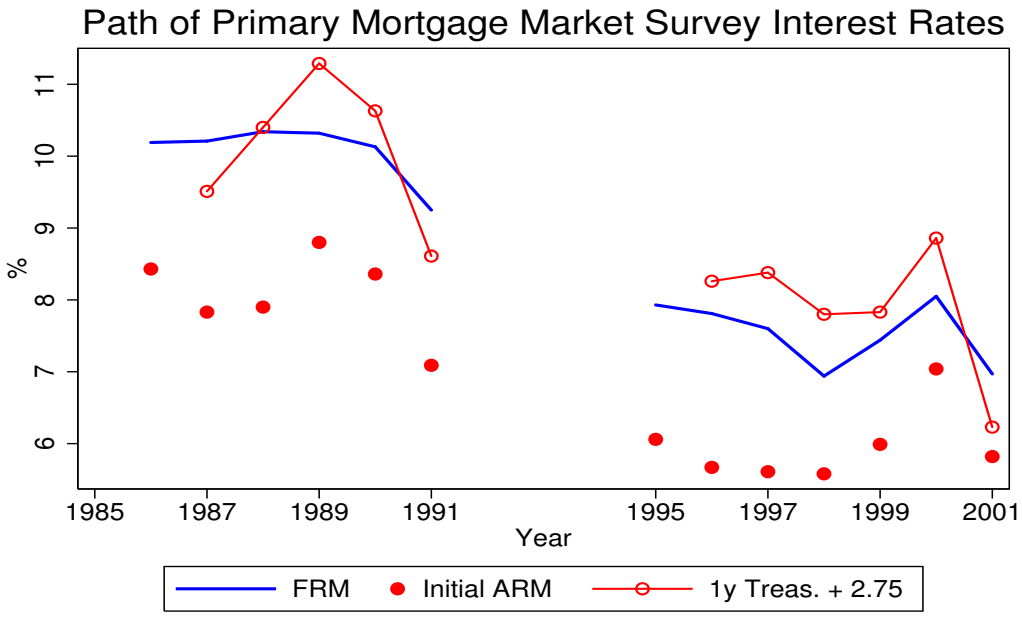
Figure 3.3



Source: 1991 & 2001 RFS / author's calculations.

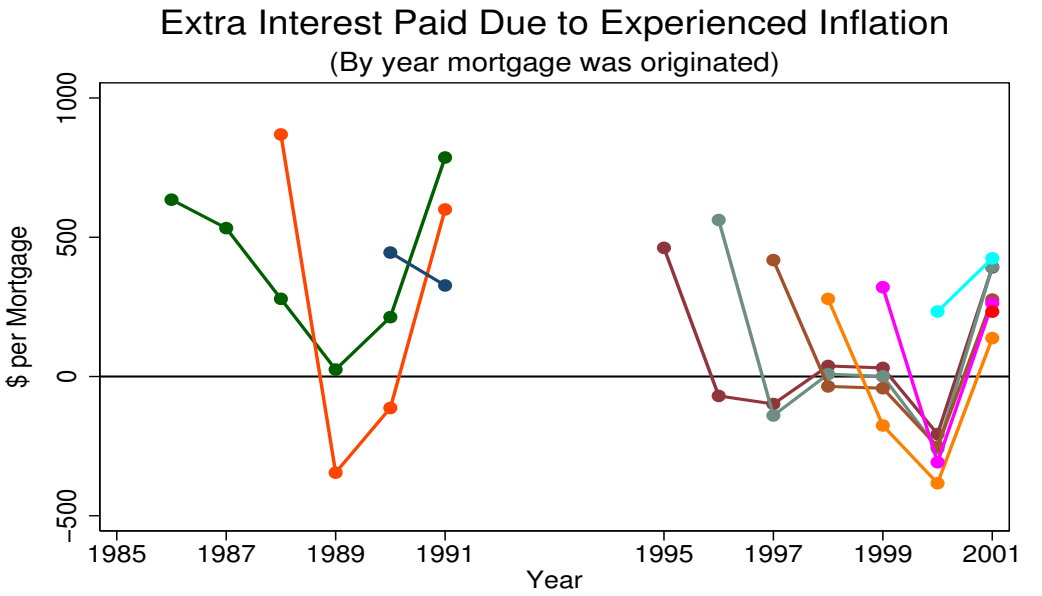
Counterfactual places a weight of zero on experienced inflation (Table 3.2, column 3).

Figure 3.4



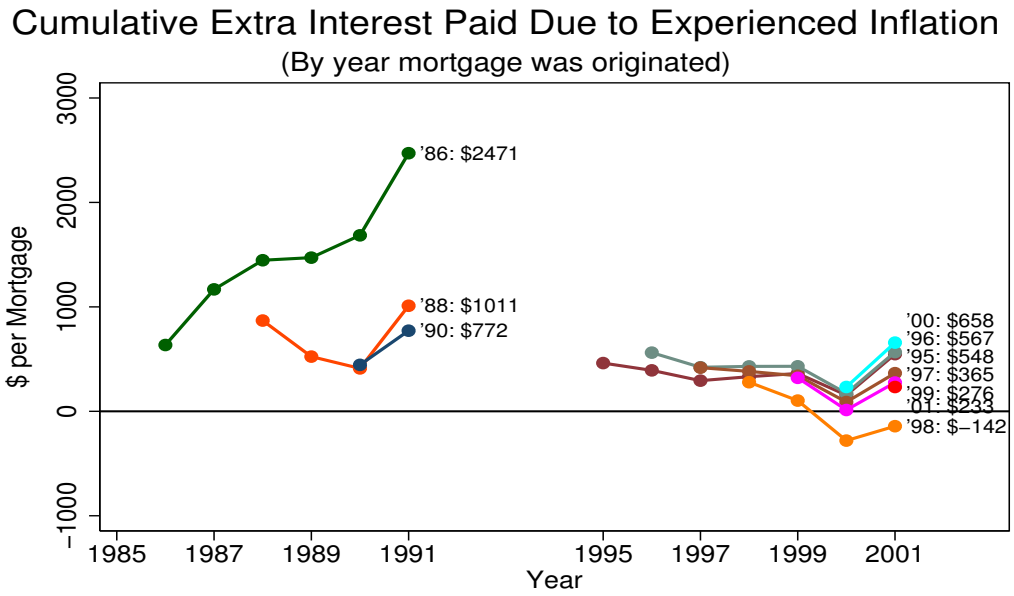
Sources: Freddie Mac, FRB.

Figure 3.5



Based on PMMS rates and 2.75 p.p. margin over 1–year Treasury, with annual resets and no caps
(Balloon mortgages omitted.)

Figure 3.6



Based on PMMS rates and 2.75 p.p. margin over 1−year Treasury, with annual resets and no caps
(Balloon mortgages omitted.)

Table 3.1: Summary Statistics

Sample of mortgages <= 6 years old at time of 1991 and 2001 Residential Finance Surveys of homeowner properties. Statistics are based on available cases. * p<0.05.

	FRM	ARM	Balloon	FRM - ARM
N=	12570	2246	734	
Contract Characteristics				
Current rate (bps)	974.8	928.8	862.6	46.0*
Initial rate (bps)	"	874.1	"	100.7*
Margin (bps)	n.a.	270.3	n.a.	n.a.
Seasoning (years)	2.6	2.8	2.1	-0.2*
Term (years)	23.4	26.3	9.0	-2.9*
Prepayment penalty?	0.060	0.094	0.053	0.0*
Economic Conditions (all in %)				
Inflation	3.38	3.47	3.61	-0.09*
FRM - ARM spread	1.76	1.86	1.69	-0.10*
Default spread	2.09	2.09	2.05	0.00
Yield spread	0.42	0.43	0.39	-0.01
Borrower Characteristics				
Primary owner age	40.5	40.9	42.0	-0.4
Experienced inflation (%)	4.77	4.81	4.68	-0.04*
Nonwhite?	0.132	0.098	0.117	0.034*
Hispanic?	0.525	0.589	0.518	-0.065*
Veteran?	0.219	0.207	0.230	0.012
Joint owners?	0.705	0.698	0.669	0.007
First-time owner?	0.434	0.371	0.364	0.063*
Has investment income?	0.283	0.302	0.249	-0.019
Has business income?	0.093	0.102	0.128	-0.010
Total income (2000 $)	71,652	80,122	69,064	-8,470*
Property Characteristics				
Central city of MSA?	0.259	0.257	0.217	0.002
Outside MSA?	0.144	0.159	0.302	-0.015
Second home?	0.012	0.018	0.016	-0.006
Mobile home?	0.034	0.020	0.045	0.014*
Condo?	0.074	0.126	0.065	-0.052*
Other Loan Characteristics				
Junior mortgage?	0.128	0.085	0.234	0.043*
Nonconventional?	0.209	0.062	0.040	0.147*
Refi?	0.254	0.246	0.297	0.008
Loan / income	1.79	2.11	1.59	-0.31*
Loan / value × 100	79.8	85.2	76.4	-5.4*
Loan / CLL	0.406	0.537	0.352	-0.131*
Jumbo loan?	0.039	0.122	0.053	-0.082*
Points paid (bps)	39.5	42.7	14.5	-3.1
Has buydown?	0.034	0.031	0.003	0.002

Notes.

Prepayment penalty clause only available for 1991. Investment income, second home status, and buydown indicator only available for 2001.

"Default spread" = Moody's seasoned corporate BAA - 10 year CM Treasury.

"Yield spread" = 30 year CM Treasury - 5 year CM Treasury.

Table 3.2: Logit Model of Mortgage Choice

Choice between FRM, Balloon, and ARM, individuals in 1991 and 2001 RFS with mortgages <= 6 years old. Omitted category for sociodemographic variables is ARM.

	(1)	(2)	(3)	(4)
Freddie Mac PMMS index rate (%)	-0.424*			
	(0.252)			
FRM Alternative-Specific Characteristics				
Freddie Mac PMMS FRM index rate (%)		-3.57***	-3.58***	-3.08***
		(0.606)	(0.606)	(0.618)
Experienced inflation in %	0.211**	0.208**	0.272***	0.239***
	(0.098)	(0.098)	(0.088)	(0.089)
Income ($ 000s)	-0.00112***	-0.00112***	-0.00112***	-0.000705**
	(0.000)	(0.000)	(0.000)	(0.000)
Age	-0.0127	-0.0125	-0.0110	-0.0060
	(0.016)	(0.016)	(0.016)	(0.017)
Age^2	0.00014	0.00013	0.00013	0.00009
	(0.00018)	(0.00018)	(0.00018)	(0.00018)
ARM Alternative-Specific Characteristics				
Freddie Mac PMMS ARM initial rate index (%)		-0.81***	-0.814***	-0.466*
		(0.266)	(0.266)	(0.270)
Balloon Mortgage Alternative-Specific Characteristics				
Experienced inflation in %	-0.2760	-0.2690		
	(0.192)	(0.192)		
Income ($ 000s)	-0.00146**	-0.00148***	-0.00149***	-0.00159***
	(0.001)	(0.001)	(0.001)	(0.001)
Age	-0.0038	-0.0044	-0.0021	-0.0336
	(0.029)	(0.029)	(0.029)	(0.029)
Age^2	0.00007	0.00008	0.00011	0.00041
	(0.00031)	(0.00031)	(0.00031)	(0.00032)
Number of Choice Situations	14,446	14,446	14,446	14,446
Log likelihood	-8443.4	-8425.2	-8426.3	-8155.2
$-\beta_{\pi, FRM} / \beta_{Rate, FRM}$	0.499	0.058**	0.076***	0.078**
(S.E. by delta method)	(0.378)	(0.029)	(0.028)	(0.033)
Alternative-specific constants	YES	YES	YES	YES
Origination year FX	YES	YES	YES	YES
Mortgage characteristics				YES

Standard errors in parentheses
*** p<0.01, ** p<0.05, * p<0.1

Bibliography

Agarwal, S., Chang, Y., and Yavas, A. (2012). Adverse Selection in Mortgage Securitization. *Journal of Financial Economics*, 105(3):640–660.

Altman, E. I. (1968). Financial Ratios, Discriminant Analysis and the Prediction of Corporate Bankruptcy. *The Journal of Finance*, 23(4):589–609.

Altman, E. I. (1984). The Success of Business Failure Prediction Models: An International Survey. *Journal of Banking & Finance*, 8(2):171–198.

Amel, D. (1993). State Laws Affecting the Geographic Expansion of Commercial Banks.

Avery, R., Brevoort, K., and Canner, G. (2007). Opportunities and Issues in Using HMDA Data. *Journal of Real Estate Research*, 29(4):351–380.

Berger, A. and Udell, G. (1995). Relationship Lending and Lines of Credit in Small Firm Finance. *Journal of Business*, 68(3):351–381.

Berger, A. N., Miller, N. H., Petersen, M. a., Rajan, R. G., and Stein, J. C. (2005). Does Function Follow Organizational Form? Evidence from the Lending Practices of Large and Small Banks. *Journal of Financial Economics*, 76(2):237–269.

Berkovec, J. and Zorn, P. (1996). How Complete Is HMDA?: HMDA Coverage of Freddie Mac Purchases. *Journal of Real Estate Research*, 11(1):39–55.

Bharath, S. and Shumway, T. (2004). Forecasting Default with the KMV-Merton Model. *AFA 2006 Boston Meetings Paper*.

Bharath, S. T. and Shumway, T. (2008). Forecasting Default with the Merton Distance to Default Model. *Review of Financial Studies*, 21(3):1339–1369.

Black, F. and Scholes, M. (1973). The Pricing of Options and Corporate Liabilities. *The Journal of Political Economy*, 81(3):637–654.

Brueckner, J. and Follain, J. (1988). The Rise and Fall of the ARM: An Econometric Analysis of Mortgage Choice. *The Review of Economics and Statistics*, 70(1):93–102.

Campbell, J. Y. and Cocco, J. F. (2003). Household Risk Management and Optimal Mortgage Choice. *The Quarterly Journal of Economics*, 118(4):1449–1494.

Carhart, M. (1997). On Persistence in Mutual Fund Performance. *The Journal of Finance*, 52(1):57–82.

Chambers, M. S., Garriga, C., and Schlagenhauf, D. (2009). The Loan Structure and Housing Tenure Decisions in an Equilibrium Model of Mortgage Choice. *Review of Economic Dynamics*, 12(3):444–468.

Dai, Z., Zhang, H. H., and Zhao, F. (2013). Tug-of-War: Incentive Alignment in Securitization and Loan Performance.

Degryse, H. and Ongena, S. (2005). Distance, Lending Relationships, and Competition. *The Journal of Finance*, 60(1):231–266.

Downing, C., Jaffee, D., and Wallace, N. (2009). Is the Market for Mortgage-Backed Securities a Market for Lemons? *Review of Financial Studies*, 22(7):2457–2494.

Fama, E. F. and French, K. R. (1993). Common Risk Factors in the Returns on Stocks and Bonds. *Journal of Financial Economics*, 33(1):3–56.

Farber, H. S. and Gibbons, R. (1996). Learning and Wage Dynamics. *The Quarterly Journal of Economics*, 111(4):1007–1047.

Frankel, D. and Jin, Y. (2011). Securitization and Lending Competition.

Garmaise, M. J. and Natividad, G. (2013). Does More Information Lead to More Financing? Local Information Shocks and Bank Credit.

Gibson, M. (1995). Can Bank Health Affect Investment? Evidence from Japan. *Journal of Business*, 68(3):281–308.

Glaeser, E. L. and Kallal, H. D. (1997). Thin Markets, Asymmetric Information, and Mortgage-Backed Securities. *Journal of Financial Intermediation*, 6(1):64–86.

Goetz, M. R., Laeven, L., and Levine, R. (2013). Identifying the Valuation Effects and Agency Costs of Corporate Diversification: Evidence from the Geographic Diversification of U.S. Banks. *Review of Financial Studies*, 26(7):1787–1823.

Green, R. and Wachter, S. (2005). The American Mortgage in Historical and International Context. *The Journal of Economic Perspectives*, 19(4):93–114.

Hahn, J., Todd, P., and van der Klaauw, W. (2001). Identification and Estimation of Treatment Effects with a Regression Discontinuity Design. *Econometrica*, 69(1):201–209.

Heckman, J. J. and Robb, R. (1985). Alternative Methods for Evaluating the Impact of Interventions: An Overview. *Journal of Econometrics*, 30(1-2):239–267.

Housing Assistance Council (2011). *What Are We Missing? HMDA Asset-Excluded Filers*.

Keys, B., Mukherjee, T., Seru, A., and Vig, V. (2010). Did Securitization Lead to Lax Screening? Evidence from Subprime Loans. *The Quarterly Journal of Economics*, 125(1):307–362.

Koijen, R. S., Hemert, O. V., and Nieuwerburgh, S. V. (2009). Mortgage Timing. *Journal of Financial Economics*, 93(2):292–324.

Kreps, D. and Wilson, R. (1982). Sequential Equilibria. *Econometrica*, 50(4):863–894.

La Cava, G. (2013). Mortgage Delinquencies and Lending Distance in the US Housing Market.

Lee, L.-F. (1978). Unionism and Wage Rates: A Simultaneous Equations Model with Qualitative and Limited Dependent Variables. *International Economic Review*, 19(2):415–433.

Liberti, J. M. and Mian, A. R. (2009). Estimating the Effect of Hierarchies on Information Use. *Review of Financial Studies*, 22(10):4057–4090.

Loutskina, E. and Strahan, P. (2009). Securitization and the Declining Impact of Bank Finance on Loan Supply: Evidence from Mortgage Originations. *The Journal of Finance*, 64(2):861–889.

Luce, R. D. and Suppes, P. (1965). Preferences, Utility, and Subjective Probability. In *Handbook of Mathematical Psychology*. Wiley, New York.

Malmendier, U. and Nagel, S. (2011). Depression Babies: Do Macroeconomic Experiences Affect Risk Taking? *The Quarterly Journal of Economics*, 126(1):373–416.

Malmendier, U. and Nagel, S. (2013). Learning from Inflation Experiences.

McCrary, J. (2008). Manipulation of the Running Variable in the Regression Discontinuity Design: A Density Test. *Journal of Econometrics*, 142(2):698–714.

McFadden, D. (1974). Conditional Logit Analysis of Qualitative Choice Behavior. In Zarembka, P., editor, *Frontiers in Econometrics*, pages 105–142. Academic Press, New York.

Merton, R. (1973). Theory of Rational Option Pricing. *The Bell Journal of Economics and Management*, 4(1):141–183.

Pearson, K. (1934). *Tables of the Incomplete Beta Function*. 1968 edition.

Petersen, M. and Rajan, R. (1994). The Benefits of Lending Relationships: Evidence from Small Business Data. *The Journal of Finance*, 49(1):3–37.

Purnanandam, A. (2011). Originate-to-Distribute Model and the Subprime Mortgage Crisis. *Review of Financial Studies*, 24(6):1881–1915.

Rubin, D. B. (1976). Inference and Missing Data. *Biometrika*, 63(3):581–592.

Salem, A. and Mount, T. (1974). A Convenient Descriptive Model of Income Distribution: The Gamma Density. *Econometrica*, 42(6):1115–1127.

Slovin, M., Sushka, M., and Polonchek, J. (1993). The Value of Bank Durability: Borrowers as Bank Stakeholders. *The Journal of Finance*, 48(1):247–266.

Stanton, R. and Wallace, N. (1998). Mortgage Choice: What's the Point? *Real Estate Economics*, 26(2):173–205.

Stein, J. (2002). Information Production and Capital Allocation: Decentralized versus Hierarchical Firms. *The Journal of Finance*, 57(5):1891–1921.

Stiroh, K. (2004). Diversification in Banking: Is Noninterest Income the Answer? *Journal of Money, Credit and Banking*, 36(5):853–882.

van der Klaauw, W. (2008). Regression-Discontinuity Analysis.

Vanasco, V. (2013). Information Acquisition vs. Liquidity in Financial Markets.

Winton, A. (1999). Don't Put All Your Eggs in One Basket? Diversification and Specialization in Lending.

Appendix A

Pure Strategy Equilibria

Theorem A.1. *Autarky.*

There is always a sequential equilibrium where no secondary market exists. In this equilibrium, the Purchaser believes that conditional on its information set being reached it faces a type 1 borrower with probability 1, so it plays $b^* = 0$. The Bank does not originate any mixed-type or bad-type mortgages. If $\phi < g(p_2R - 1)$, then the Bank plays $a^* = 1$ and originates good-type borrowers. If $\phi > g(p_2R - 1)$, then the Bank plays $a^* = 0$ and no mortgages are originated. Otherwise the Bank mixes.

That is, if audit costs ϕ are relatively small, then the Bank will audit, originate, and hold good types in autarky. If audit costs are big, the Bank will prefer holding risk-free bonds to extending loans.

Theorem A.1 may be read in terms of lender-borrower distance. Supposing that audit costs do increase with distance, this theorem describes a situation where all bank lending is local. A potential borrower living in a remote or rural area without any local banks will be excluded from bank credit markets, not because she cannot repay but because it is too costly for the Bank to learn her type and separate good borrowers from bad borrowers. Secondary markets overcome this market failure by allowing investors to diversify and hold both types.

Theorem A.2. *Screening Costs as a Signal.*

If $\phi \geq g(p_2R - 1 - \beta\theta)$, then a sequential equilibrium exists in which the Bank plays $a^* = 0$ and the Purchaser plays $b^* = 1$, with the remainder of the Bank's strategy given by Lemmas 2.1, 2.2, and 2.3. The Bank originates and sells the loans of all potential applicants without screening them, and the Purchaser buys for sure.

Recall that case (i) of Theorem 2.1 describes an equilibrium wherein screening costs ϕ and the Purchaser's equilibrium buy strategy b^* rise together. This Theorem states that if ϕ is large enough that it is not credible for the Bank to ever want to pay, then the mere fact that such a mortgage is being offered for sale reveals that the borrower was not screened. The intuition is that if it is sufficiently costly for the Bank to acquire information, and if this cost is observable to the Purchaser, then the Purchaser should believe that the Bank does not have any private information.

These results can be interpreted in terms of a signaling game. If the Bank originates a low screening-cost borrower, the Purchaser should believe that the Bank has screened this borrower for sure, so any mortgage being offered for sale has been adversely selected and is a bad type. No secondary market can operate under these conditions, and all mortgage funding must come internally from the Bank. As search costs increase, it becomes more credible that the Bank has not screened. Theorem A.1 describes a cutoff for ϕ above which it is not profitable for the Bank to screen when there is no chance it can sell the loan. No mortgage lending will occur for borrowers who are too costly to screen and the market will break down. Theorem A.2 describes a cutoff for ϕ above which it is not profitable for the Bank to screen even when there is a 100% chance of selling the loan. This cutoff is smaller than the

cutoff given in Theorem A.1 because the benefits of not screening are larger when the Bank can sell loans on a secondary market and earn its share of gains to trade θ. This means that there is a region of auditing costs ϕ between the two cutoffs where the Bank could hold either set of beliefs about the Purchaser's strategy and play either strategy. Offering to sell a mortgage which would have cost the Bank $g(p_2R - 1 - \beta\theta) < \phi < g(p_2R - 1)$ to audit does not clearly signal which strategy the Bank is playing.

A further comment about this region of indeterminacy is in order. Note that the Bank's profits are greater in the secondary market equilibrium than in the autarky equilibrium. Forward induction would suggest that this is the "correct" equilibrium. Of course, the Bank might be unsure whether the Purchaser realizes that the Bank should behave according to this equilibrium (and vice versa). A mixed strategy equilibrium might provide the most appealing set of beliefs, given this mutual uncertainty. This would return us to the world of Theorem 2.1 case (ii).

Appendix B

Proofs

Proof of Lemma 2.1. Consider a history where the Bank has audited and originated a type 2 borrower. Its expected payoff from holding the loan is $p_2R - \phi$ and from attempting to sell the loan is $b(1+\beta\theta-\phi)+(1-b)(p_2R-\phi)$. Auditing costs are sunk, and by Assumption 2, $p_2R > 1 + \beta\theta$, so for any set of beliefs with $b > 0$, the Bank is strictly better off playing $s_2^* = 0$. The proof is similar for type 1 borrowers and unaudited types. However, the Bank's strategy is not pinned down if it believes that $b = 0$, since the sequential equilibrium concept does not rule out weakly-dominated strategies. □

Proof of Lemma 2.2. By Assumption 2, $p_2R > 1$, so the Bank will play $o_2^* = 1$. □

Proof of Lemma 2.3. **(a)** By Lemma 2.1, the Bank will play $s^* = 1$ if it reaches that decision node. So in histories where it does not audit, the Bank will strictly prefer originating to denying the borrower's loan application if $b(1+\beta\theta)+(1-b)(mR-\kappa) > 1$. Since $mR = 1$, this simplifies to $b > \kappa/(\beta\theta+\kappa)$.

(b) By Lemma 2.1, the Bank will play $s_1^* = 1$ if it reaches that decision node. So in histories where it audits and nature selects a type 1 borrower, the Bank will strictly prefer originating to denying the borrower's loan application if $b(1+\beta\theta)+(1-b)(p_1R) > 1$. This simplifies to $b > (1-p_1R))/(1+\beta\theta-p_1R)$. □

Proof of Corollary 2.1. The two cutoffs given in Lemma 2.3 are of the form $x/(\beta\theta+x)$, which is increasing in x. From Assumptions 3 and 4, we have that $\kappa < \theta < 1-p_1R$. This leads to the result that the cutoff for originating mixed-type mortgages in part (a) is below the cutoff for originating type 1 borrower in part (b). □

Proof of Theorem 2.1. Given that the Purchaser believes the Bank is also playing $s^* = 1$, $s_1^* = 1$, $s_2^* = 0$, and $o^* = o_1^* = o_2^* = 1$, equation 2.3 gives the beliefs the Purchaser must hold about the Bank's audit strategy which leave it indifferent between buying and not buying:

$$a^*[g(1)+(1-g)(p_1R+(1-\beta)\theta]+(1-a^*)[mR-\kappa+(1-\beta)\theta]=1$$

Given the remainder of the Bank's strategy, Equation 2.4 is the set of beliefs the Bank must hold about the Purchaser's buy strategy which leave it indifferent between auditing and not auditing:

$$g(p_2R-\phi)+(1-g)[b^*(1+\beta\theta-\phi)+(1-b^*)(p_1R-\phi)]=b^*(1+\beta\theta)+(1-b^*)(mR-\kappa)$$

Both simplifications rely on Assumption 1 and $mR = 1$.

Assumptions 2, 4, and 5 guarantee that $a^* < 1$. Either condition (i) or (ii) is necessary for $b^* \leq 1$. □

Proof of Corollary 2.2. **(a)** $g[p_2R - 1 + (1-\beta)\theta] > g(2-\beta)\theta > (1-\beta)\theta$ by Assumptions 2 and 5, so both numerator and denominator of the fraction are positive.

(b) Under the stated condition, the numerator of the fraction simplifies to $(\phi - \kappa)(c_\phi + c_\kappa)$ and the result follows.

□

Proof of Theorem A.1. Consider the case where $\phi < g(p_2R - 1)$. We must specify a sequence of fully-mixed strategies over the Bank's actions consistent with the limiting strategy $\{a = 1,\ o_2 = 1,\ s_2 \geq 0,\ o_1 = 0,\ s_1 \leq 1,\ o = 0,\ s \leq 1\}$. Let $\{a_n = 1 - \varepsilon,\ o_{2,n} = 1 - \varepsilon,\ s_{2,n} = \varepsilon^2/(1-\varepsilon),\ o_{1,n} = \varepsilon + \varepsilon^2,\ s_{1,n} = \varepsilon/(\varepsilon + \varepsilon^2),\ o_n = \varepsilon + \varepsilon^2,\ s_n = \varepsilon/(\varepsilon + \varepsilon^2)\}$. By Bayes' Rule the Purchaser will believe $\mu_n(Bad|S) = (1-g)(\varepsilon - \varepsilon^2)/[(1-g)(\varepsilon - \varepsilon^2) + g(\varepsilon^2 - \varepsilon^3) + \varepsilon^2] \to 1$ conditional on its information set being reached, so its best response at the limit is $b^* = 0$. The Bank's limiting strategy is a best response to $b^* = 0$ by Lemmas 2.1, 2.2, 2.3, and the fact that its expected profits from auditing and originating exceed those of not auditing and not originating: $g(p_2R) + (1-g)(1) - \phi > 1$. A similar construction exists for $\phi > g(p_2R - 1)$. □

Proof of Theorem A.2. This is a generalization of case (i) of Theorem 2.1. If $\kappa < g(p_2R - 1 - \beta\theta) \leq \phi$, then the Bank prefers not auditing for any value of $b \in [0, 1]$. The preference is strict if the inequality is strict. The Bank's strategy $\{s = 1,\ s_1 = 1,\ s_2 = 0,\ o = o_1 = o_2 = 1,\ a = 0\}$ and the Purchaser's strategy $b = 1$ are sequentially rational and the associated generated beliefs are consistent. □

Appendix C

Algorithm for Estimating County Income Distributions

The Census Bureau's Small Area Income and Poverty Estimates program provides estimates for the median household income, $y_{0.5}$, and the fraction of people in poverty, $PctPov$, for counties and states in the years 1989, 1993, 1995, then annually beginning in 1997. I fill in the missing years by linear interpolation of the poverty rate and log-linear interpolation of median household income. The U.S. poverty threshold is defined nationally but differs by family size, age of householder, and number of children. The Census Bureau reports historical weighted average poverty thresholds for families of different sizes since 1959, where the weights are by the presence and number of children. I define y_{pov} as the "all ages" weighted average poverty threshold for a household with 2.5 people (average of 2 and 3 people) in each year.

Taking a parametric approach, I assume that the income distribution in every county-year follows a gamma distribution, where $\alpha > 0$ is the "shape" parameter and $\beta > 0$ is the "scale" parameter. This produces a system of two equations in two unknown parameters along the gamma CDF:

$$\Gamma_{\alpha,\beta}(y_{0.5}) = 0.5$$

$$\Gamma_{\alpha,\beta}(y_{pov}) = PctPov$$

I use a numerical solver from the R package **rootSolve**[1] to solve the system and estimate a different gamma income distribution for every county-year. The major challenge is that this system is not globally concave, so a Newton-Raphson solver only converges when starting from parameter values near a root. I use the following algorithm to locate "good" starting values.

1. Find an approximate solution to use as a starting value for the full numerical solver. Call the output of this step (α_0, β_0).

   (a) Concentrate the CDF on the shape parameter α using the Salem and Mount (1974) formula: $\beta(\alpha) \approx 3y_{.5}/(3\alpha - 1)$, and search for α_0 which is a root of the poverty quantile equation[2]

   $$Q(\alpha) = (\chi^2_{(2\alpha)})^{-1}(PctPov) - \frac{2}{\beta(\alpha)} y_{pov}$$

   (b) Search for a root $\alpha_0 \in [1, 20]$. If successful, return $(\alpha_0, \beta(\alpha_0))$.

   (c) If no solution is found, search for a root $\alpha_0 \in [0.34, 1]$. If multiple solutions are found, take the largest one: return $(\max(\alpha_0), \beta(\max(\alpha_0)))$.

   (d) If no solution is found, return $\alpha_0 = 0.15$ and $\beta_0 = \beta(0.35)$.

2. Use a Newton-Raphson numerical solver to locate an exact solution to the two-equation system. Call the output of this step (α_1, β_1).

[1] Soetaert (2010), version 1.6.4.

[2] If $Y \sim \Gamma(\alpha, \beta)$, then $2/\beta \cdot Y \sim \chi^2_{(2\alpha)}$.

(a) Use (α_0, β_0) as initial values in the solver.

(b) If the NR solver does not converge, sequentially try the starting values $(a, \beta(\max(a, 0.35))$ from a grid of values $a \in \{0.05, 0.10, ..., 1, 2, ..., 100, 200, ..., 1000\}$.

3. To handle extreme parameter values and non-convergence in Step 2:

(a) Winsorize the top and bottom 1% of values of α_1 across counties within every year. For non-converged systems: county-years with $PctPov$ equal to zero are top-coded; those where the fraction is ≥ 0.50 are bottom-coded. Call the Winsorized output α_2.

(b) Apply a numerical solver to locate the value β_2 solving the median equation: $\Gamma_{\alpha_2,\beta_2}(y_{0.5}) = 0.5$, starting the search at $(\alpha_2, \beta(\alpha_2))$.

The same procedure may be applied to estimate state income distributions. In all state-years, an exact solution is found involving $\alpha > 1$.

Salem and Mount (1974) derive the following expression for the Gini concentration index of a gamma distribution:

$$G_\gamma = 2 \cdot I_{0.5}(\alpha, \alpha + 1) - 1$$

where $I_{0.5}(a, b)$ is the CDF of the beta distribution, referred to by K. Pearson (1934) as the "incomplete beta-function ratio,", evaluated at $x = 0.5$.