Price Discrimination and Adverse Selection in the U.S. Mortgage Market *

Huihan Zhang †

October 31, 2024

Click Here for the Latest Version.

Abstract

Interest rates are highly dispersed in the U.S. mortgage market, even for loans from the same lender to observably similar borrowers. This paper studies the source of conditional price dispersion by estimating a structural model that features borrowers' demand for mortgages and lenders' individualized pricing strategies in a market characterized by adverse selection. Cost-based risk adjustment contributes more to conditional price dispersion than demand-based price discrimination. Equilibrium interest rates increase if lenders must set uniform interest rates for similar borrowers. I provide suggestive evidence that lenders' ability to learn borrowers' private information and tailor interest rates can alleviate adverse selection.

Keywords: Price Dispersion, Adverse Selection, Home Ownership, Regulatory Policies

JEL code: D43, D82, G21, G51, L11, L51

*I am deeply indebted to my advisor Marc Rysman for invaluable guidance and constant support, and to Jihye Jeon, and Andrey Fradkin for advice and encouragement. I also thank Rubaiyat Alam, Alexei Alexandrov, Randall Ellis, Adam Guren, Jean-François Houde, Erica Jiang, Chuqing Jin, Tim Lee, Siqi Li, Yan Liu, William Matcham, Pasquale Schiraldi, and participants at the BU empirical micro seminar, CES, SGE, and AMES for helpful comments and discussions. CES, AMES, and BU IED provided travel grants. All errors are my own. I declare that I have no relevant or material financial interests that relate to the research described in this paper.

†Amazon. Email: hzhangecon@gmail.com.

1 Introduction

Recent research has shown substantial price dispersion in the U.S. mortgage market, with lenders charging similar borrowers different rates. This contrasts with the U.K. mortgage market, where mortgage rates are much less consumer-based (Robles-Garcia, 2019). Borrowers with the same characteristics can get rates of 0.7 p.p. apart, which translates to over \$200 per month for the median mortgage loan. Understanding mortgage rate dispersion is important because of the mortgage market's central role in the U.S. consumer credit market. As of 2023, there are about \$12 trillion in outstanding mortgage loans in the U.S. Moreover, mortgage loans represent the most significant borrowing decision households make in their lives. Hence, efficient pricing of mortgage rates has important welfare implications for U.S. households.

In this paper, I quantify market outcomes and welfare implications of price dispersion and adverse selection in the U.S. mortgage market. I study two channels that could explain the observed price dispersion. Firstly, suppose lenders have market power and can learn about borrowers' willingness to pay (WTP). In that case, they may price discriminate against the borrowers, i.e., charging higher prices to borrowers with higher WTP. Secondly, lenders can learn about borrowers' private risks and adjust prices for risks. Distinguishing the two channels is important because they have different policy implications. For the first channel, lenders' goal of price discrimination is to extract consumer surplus. On the other hand, individual-level risk-based price adjustment under the second channel may improve total welfare: low-risk borrowers who were not in the market due to high unique prices may now be served with individualized pricing. If the first channel dominates, implementing price regulations that reduce price dispersion may enhance borrowers' welfare. Alternatively, if the efficiency gains from cost adjustments dominate, then consumers may be better off from the status quo.

Moreover, I study the interaction between price discrimination and adverse selection. Adverse selection characterizes the market if more costly borrowers are more likely to demand mortgage loans. It is known that adverse selection negatively impacts market outcomes, leading to higher prices and less lending. However, the question of whether interest rates are more widely dispersed as a result of adverse selection has been studied less. Conversely, how tailoring interest rates affects adverse selection is theoretically ambiguous as it depends on the competitive environment. I show that lenders' ability to learn information about borrowers' private risks to tailor individualized rates reduces the extent of adverse selection in the U.S. mortgage market.

I use two administrative data sets. The first is the Home Mortgage Disclosure Act (HMDA) data set, which contains loan-level data from the near-universe of U.S. mortgage loans. The data spans from 2007 to 2021 and measures lenders' identity, detailed loan and borrower characteristics, and the location where the loan originated at the census tract level. The data after 2018 also contains price measures, including interest rates and non-interest fees, i.e., discount points and lenders' credits. The second data set is the Fannie Mae and Freddie Mac Single Family Loan Acquisition data set, which contains a subset of loans sold to government-sponsored entities (GSEs). The GSEs data provided additional loan information, including borrowers' credit scores and loan default behavior.

I first present new facts about the interest rate dispersion in the U.S. mortgage market. I show that there is substantial residual interest rate dispersion after controlling for lender, state, borrower, and loan characteristics. A variance decomposition analysis suggests that lender identity accounts for 20% - 30%, state and year explain 15% - 20%, loan and borrower characteristics explain 5% of the variation in interest rates, and the residual 50% variation is unexplained. I verified using the GSE data that borrowers' credit score only accounts for about 10% of price dispersion. Therefore, about 40% of the dispersion cannot be explained by lenders' identity or observed risk measures. I also describe two possible channels to explain this residual price dispersion. Leveraging on market-level variation in the extent of lenders' market power, I show that price dispersion is not merely a result of risk-based price adjustment. In particular, in markets where lenders have greater power, price discrimination based on WTP is more prominent, resulting in greater price dispersion.

Moreover, I find empirical evidence consistent with adverse selection in the mortgage market. Adverse selection implies that borrowers who are willing to accept higher prices are also likely to be unobservably riskier. As I do not observe *ex-post* loan defaults, I use lenders' loan denial decisions to proxy borrowers' riskiness. If the borrowers are also more likely to be rejected by lenders, it suggests that lenders are cautious about extending credit to individuals they view as risky. An additional challenge is that I do not observe interest rates for loan applications that are denied. To address this challenge, I exploit a discontinuity in interest rates around the loan-to-value (LTV) cutoffs (Tsai, 2023) and use this discontinuity in interest rates to study application denials at the cutoffs. It has been documented that mortgage prices are impacted by the GSE's price-setting rules, which follow a step function in LTV and credit scores (Agarwal et al., 2018). Hence, the pricing rule generates a quasi-random variation in interest rates for borrowers around the cutoffs. I find that both interest rates and rejection rates jump up discontinuously at the LTV cutoffs, consistent with adverse selection in the market.

Motivated by the reduced-form evidence, I develop and estimate a structural model of the mortgage market that allows me to estimate the extent of adverse selection and run counterfactual scenarios that are of concern to policymakers. The model distinguishes three sets of information: (1) information known by everyone (public), (2) information known by borrowers and lenders but not to the researchers (soft), as well as (3) information known only by borrowers (private). Lenders offer mortgage loan contracts to borrowers with interest rates and non-interest costs. Lenders are differentiated by the type of institution, i.e., banks vs non-banks, and the number of years they have been in a market. Borrowers seek mortgage loans to finance a house, the riskiness of whom is their private information. Borrowers choose a lender from which to borrow or the outside option according to a random coefficient logit demand model. Lenders face three types of information: hard information (observable covariates), soft information (observed by lenders but not to the researchers), and private information (unobserved by lenders). I refer to soft information as signals that lenders can observe and use to learn about borrowers' private information, and I classify private information as demand and cost shocks separately. Lenders face uncertainty both in borrowers' private information on demand and in the costs of lending to the borrower. Thus, they compete 'a la Bertrand-Nash on expected individualized interest rates over the two types of shocks. There are two critical correlations in the model: that between the unobservable determinants of demand and costs and that between lenders' signals and the unobservable shocks. When the first correlation is positive, I infer it as the market is characterized by adverse selection: riskier borrowers are more likely to demand a mortgage. When the second correlation is high, I say that lenders' signals are effective in learning about borrowers' private types.

My model is similar to Crawford et al. (2018) but differs in the specification of the correlation structure and the assumptions of demand signals. First, a novel component of my model is that it does not require loan performance data. Instead, I model adverse selection as a correlation between unobserved demand and cost shocks. Second, my model does not require panel data with repeated borrowing for the same borrower, which is often not available to the public in the U.S. mortgage market. To overcome this data limitation, I assume that the proxies for price signals are random effects instead of fixed effects.

I estimate the model on the originated loans from 2018 to 2022. My estimation proceeds in three steps. Borrowers' choices and lenders' optimal pricing equation jointly pin down the variances and covariances among the distributional parameters. With the assumption that the variation in market structure at the market level is exogenous to the distribution of unobserved risks across borrowers, the model can separately identify the structural parameters for the price discrimination and risk adjustment channel. Lastly, I exploit the variation in costs across markets (a.k.a. Hausman IVs) - the average interest rates in other markets - to identify borrowers' price elasticities. Lastly, with the estimated demand parameters, I estimate the supply parameters that break down lenders' optimal pricing to marginal costs and markups. Using the estimated covariance matrix, I structurally decompose the conditional price dispersion into risk-based and non-risk-based components.

In my results, I find evidence consistent with adverse selection. The estimated standard deviation of the unobserved determinants of demand and cost are 0.53 and 0.58, respectively. The estimated correlation is 0.43 - the more unobservably costly types are also those that have higher demand for mortgages. Moreover, the correlation parameter conditional on lenders' signals about the borrowers is 0.30. This suggests that the ability to price based on lenders' soft information alleviates the extent of adverse selection in the mortgage market. Moreover, lenders learn borrowers' private types through signals. I estimate that the effectiveness of lenders' signals[1] for borrowers' private demand types is 0.13 and for borrowers' private cost types is 0.91. This suggests that lenders' signals that they use to set prices are more effective for hidden cost information and less so for borrowers' demand. Consequently, the conditional price dispersion in the mortgage market is more from lenders'

[1]The effectiveness of lenders' signals is defined as the correlation between signals and lenders' private types

cost adjustments by learning about borrowers' private risks rather than price discrimination by learning about borrowers' demand.

I run counterfactual simulations to quantify the effects of price dispersion and adverse selection. In the first counterfactual, I impose restrictions on price dispersion and analyze their effects on different market structures and types of borrowers when the market re-equilibrates. I do so by limiting lenders' price-setting ability: lenders can only set a uniform price for borrowers who share the same observable covariates. I find that limiting individualized pricing conditional on observable covariates increases the equilibrium interest rate by 1.6% on average and decreases consumer surplus by 18.6%. The effects are heterogeneous across borrower types. The top-income quintile borrower experienced the largest increase in equilibrium prices, suggesting that the top-income borrowers benefit the most from individualized pricing. For the supply side, market concentration increases and lenders' profits increase overall compared to the status quo. In sum, the results suggest that consumers benefit from a certain degree of individualized pricing because it lowers equilibrium prices through more intense concentration. I offer two interpretations of the counterfactual results. First, competition reduces as a result of uniform pricing. Lenders now need to lower prices for all borrowers in a group to undercut competitors. Thus, equilibrium interest rates increase as a result of reduced competition in the market. Second, adverse selection worsens when lenders cannot set prices based on signals about private risks. Adverse selection typically leads to higher prices under imperfect competition (Crawford et al., 2018). Hence, when lenders cannot price discriminate within a group, interest rates increase as a result of worsened adverse selection. Distinguishing the two channels is beyond the scope of this paper and would be worth studying for future research.

The paper proceeds as follows. Following the literature review in section 2, I describe my data in section 3 and present my reduced-form findings in section 4. An exposition of my structural model follows in section 5. In section 6, I explain model estimation. Section 7 presents my demand and supply parameter estimates, and in section 8, I describe the results of the counterfactual analyses. Section 9 concludes the paper.

2 Contribution to the Literature

This paper establishes a link between the finance literature on selection problems and the price discrimination literature on competition and market power. First, this paper contributes to the literature on explaining price dispersion in general (Borenstein and Rose, 1994) and specifically in the U.S. mortgage market (Bhutta et al., 2020; Alexandrov and Koulayev, 2018). Existing research emphasizes rate dispersion by different lenders among observably similar borrowers. Due to search costs and low financial sophistication, borrowers may not always choose the best available offer, leading to high rate dispersion (Bhutta et al., 2020). This paper endogenizes the supply side of mortgage loans and studies the equilibrium outcomes of changes in price dispersion.

Second, it contributes to the voluminous literature that measures the importance of adverse selection in lending markets. Einav et al. (2021) surveys this literature. Most existing studies assume that the prices are uniform across borrowers conditional on observable characteristics. This paper allows for individualized pricing based on unobservables, therefore, incorporating the interaction between lenders' price discrimination decisions with adverse selection in the market. More closely related to my paper include Crawford et al. (2018) which allows individual-specific pricing. It does not distinguish the source of price dispersion which is the main conceptual contribution of this paper. I show that a decomposition of price dispersion has important policy implications. Restricting the extent of residual price dispersion may improve consumer surplus if it is largely from price discrimination.

Lastly, this paper is related to an emerging literature on industrial organizations in financial markets. Clark et al. (2021) reviews this literature. Recent literature includes studies on the business lending market (Crawford et al., 2018), credit card market (Nelson, 2018; Matcham, 2022), security analyst market (Jin, 2022), and mortgage market (Allen et al., 2019; Robles-Garcia, 2019; Benetton, 2021; Tsai, 2023).

3 Industry Background and Data

The residential mortgage market is the largest consumer credit market in the U.S. As of 2023, there has been an outstanding debt of about \$14 trillion. I focus my analysis on mortgage loan origination, the process by which a borrower secures a mortgage.

I begin by describing the U.S. mortgage market: loan types, government involvements, and non-interest-rate fees in the U.S. mortgage market. Then I discuss two main sources of variation in mortgage interest rates for similar borrowers and loans: market power and unobserved (to us) risks.

3.1 U.S. Mortgage Market

Loan Types I focus my analysis on two main residential mortgage market segments: the conforming loan market and the jumbo loan market. Together, These two segments account for about 80% of total mortgage loans, with conforming loans accounting for over 70%. Conforming loans must be below the conforming loan limit, which was $417K before 2017 and has grown since then to $726K in 2023. Jumbo loans refer to mortgages that exceed the conforming limits are called. I exclude the Federal Housing Administration (FHA) and Veterans Administration (VA) loans in my analysis.

Government Involvement The government is highly involved in the U.S. mortgage market. Government-sponsored enterprises (GSEs) purchase conforming loans from mortgage lenders and offer guarantees to loan defaults. The sold loans are then packaged into mortgage-backed securities (MBS) and resold to investors. According to the Home Mortgage Disclosure Act (HMDA), Fannie Mae and Freddie Mac secured about 50% of total originated residential mortgages in the U.S. Loans that exceed the conforming loan limit cannot be secured by the GSEs.

Non-interest Fees In the U.S., Non-interest fees include origination charges (a form of underwriting costs), discount points, and lender credits. Discount points are a form of upfront costs that borrowers can purchase to lower interest rates. Mortgage discount points typically cost 1% of the principal and lower interest rate by 0.25 p.p.[2] Lender credits are the opposite of discount points where borrowers can take some credit from the lenders and in turn get higher interest rates. In my analysis, I compute effective interest rates taking into account discount points and lender credits.

[2] For example, if a mortgage is $200,000 and the interest rate is 4.5 percent, one point costs $2,000 and lowers the monthly interest to 4.25 percent.

3.2 Price Dispersion

Mortgage interest rates in the U.S. are highly dispersed, with each lender charging similar (to us) borrowers different prices for similar loans.[3] I explain this price dispersion among similar borrowers and loans by two sources: market power and unobserved risks.

Source of Market Power Many borrowers shop infrequently for mortgages. According to the National Survey of Mortgage Origination (NSMO), most borrowers believe they would get the same rate from any lender. Many borrowers only seriously consider one lender; some consider two, and barely any consider more than two. These survey responses provide initial evidence that search frictions are important in this market. Borrowers' infrequent searches leave room for the exercise of lenders' market power.

Source of Risks Two major sources of risks are default risks and prepayment risks. Default risks are major for jumbo loans, whereas the majority of conforming loans are sold to the GSEs and hence are insured for default risks. Another source of lenders' risks is prepayment risks. After loan origination, lenders may choose to retain loan servicing rights. Hence, lenders lose revenue if the loans are paid prematurely. The main driver of observed price dispersion is any information discrepancy between us and the borrowers and lenders. If lenders learn more about borrowers' risks than we do, we will observe price dispersion conditional on all hard information about risks. Since my study focuses only on loan origination, I specify a model that is general about types of risks to lenders. I assume that lenders account for all types of risks that are potential costs to them at origination.

3.3 Data

I use two public data sources. The first data set is the Home Mortgage Disclosure Act (HMDA), enacted by Congress in 1975 by the Federal Reserve Board's Regulation C to determine whether lenders are meeting communities' housing needs and to monitor any discriminatory lending patterns. In 2011, the rule-writing authority was transferred to the Consumer Financial Protection

[3] This contrasts with the U.K. mortgage market where mortgage rates are not customer-based (Robles-Garcia, 2019; Benetton, 2021).

Board (CFPB). This regulation applies to certain financial institutions, including banks, savings associations, credit unions, and other mortgage lending institutions (Federal Financial Institutions Examination Council, 2021). The data set covers near-universe US home loans[4]. Public HMDA data covers from 2007 to 2022. HMDA 2007-2017 reports lenders' identity, loan amount, rate spread, applicants' income, race, loan purpose, loan type, and occupancy type. The rate spread is the amount above the average prime offer rate (APOR). HMDA records the rate spread if it is 3 p.p more than APOR before 2009 and if 1.5 p.p more than APOR in 2009-2017[5] (Banga, 2022). In 2018, as part of implementing the Dodd-Frank Act, the CFPB required additional information to be reported - HMDA after 2018 reports interest rate, loan-to-value ratio, debt-to-income ratio, property value, applicants' age, and fixed fees at loan closure. In my study, I focus on 30-year-fixed-rate loans that are originated for purchases of homes that are not for commercial purposes. I also exclude FHA and VA loans from my study. For structural estimation, I further exclude small lenders that have smaller than 2% market shares.

The second data source is Fannie Mae and Freddie Mac's Single Family Loan Performance data set. Data is available from 2000Q1 to 2023Q1. According to HMDA, about 50% of loans on average are sold to GSEs between 2007 and 2017. I observe lenders' identity [6], loan amount (unpaid principal balance), origination channel (broker, retail, correspondent), interest rate, loan-to-value ratio (LTV), debt-to-income ratio (DTI), loan purpose (Cash-out refinance, refinance, purchase), borrowers' credit score, property type (i.e. Condo, Co-op, PUD, manufactured home, single-family home), and property zip code.

Table 1 shows the summary statistics for the key variables for the main studied period 2018-2022. The average mortgage loan rate is 3.92 percentage points and the lenders charge on average an origination fee of \$1.8K. The average discount point paid upfront is about \$2K. The average mortgage loan size is \$332K with a loan-to-value ratio of 82%. The average borrower has an annual household income of \$135K. There are more than 200 lenders per

[4]Only small lenders that do not meet the reporting criteria are exempt. https://www.ffiec.gov/hmda/pdf/DepCriteria0204.pdf

[5]This results in about 95% missing rate spread in 2007-2017.

[6]For sellers that represent less than one percent of volume within a given acquisition quarter as represented by the original unpaid principal balance, "Other" will be displayed in this field.

Table 1: Summary Statistics

	Obs (m)	Mean	St. Dev.	Min	Max
Interest Rate (p.p.)	10.9	3.92	1.12	0.001	8.50
Income (k$)	10.9	135	114	24	724
Loan Amount (k$)	10.9	332	2420	5	1,445
Loan-To-Value (%)	10.9	82	15	27	102
Nonbank (%)	10.9	63	48	0	100
Conforming Loans (%)	10.9	93	26	0	100

market in the raw data and the number of lenders drops to 12 per market after I exclude the small lenders. .

Figure 1: Evolution of Mortgage Rates and Dispersion

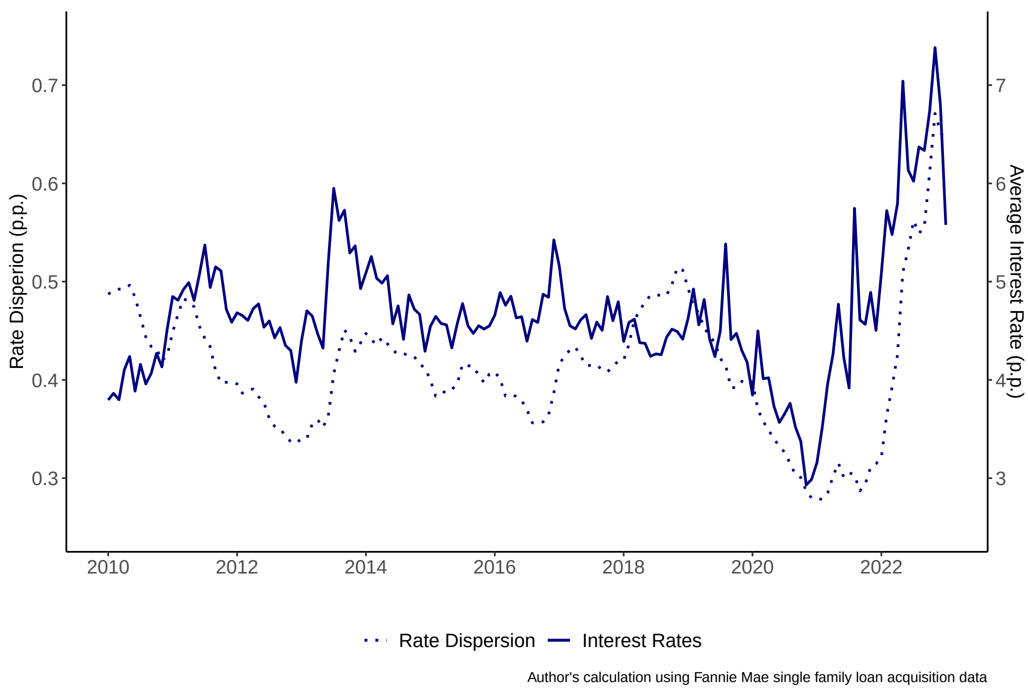


Figure 1 shows monthly average mortgage rates and rate dispersion measured in standard deviation. Both mortgage rates and rate dispersion have risen substantially since 2021, reaching the highest level in the past decade in 2023. The nominal standard deviation reached 0.7 p.p. in 2023 which translates to more than $90K in total or over $200 per month for the median mortgage loan.

4 Reduced-Form Evidence

Mortgage rates are determined by market, lender, borrower, and loan characteristics. In this section, I conduct a variance decomposition analysis to show how much each category contributes to the observed price dispersion. I show that substantial dispersion remains even after we control for these characteristics. The conditional price dispersion can be explained by lenders' price discrimination and risk adjustments. On the one hand, lenders may extract surplus from borrowers by charging up to their willingness to pay. On the other hand, lenders may adjust prices for risks if they learn about borrowers' private risk information. I present motivating evidence suggesting that both channels are present in the US mortgage market.

4.1 Variance Decomposition

To understand how much each factor contributes to price dispersion, I estimate the following regression at the loan level including a set of fixed effects (inc. discretized continuous control variables)

$$r_i = \alpha + \sum_{x=1}^{4} \sum_{j=1}^{J_x} \beta_{xj} \mathcal{I}\{x_i = j\} + \varepsilon_i$$

where x refers to characteristics - lender, state, demographics characteristics (income and age), and loan characteristics (loan-to-value ratio and loan amount), and j refers to fixed effects dummies or discretized categories for continuous variables.

I compute the group-level standard deviation sd_x by taking $sd(\beta_{xj})$'s across entities/levels j. The share of standard deviation attributed to characteristics x is defined as

$$sd\_share_x = \frac{sd_x}{\sum_x sd_x + sd(\varepsilon_i)}$$

Table 2 presents the results of the variance decomposition. Lender and state account for about 30% of the variance in price and loan, and borrowers' demographics explain about 15% of price variation. 55% variance remains unexplained, suggesting similar borrowers are charged different prices for the same loan size and loan-to-value ratio. This dispersion greatly impacts household wealth, particularly in low- to moderate-income households. Understanding the cause of conditional price dispersion is important for home ownership

Table 2: Variance Decomposition

	Variance Decomposition				
	2018	2019	2020	2021	2022
Lender	0.26	0.23	0.25	0.27	0.20
State	0.08	0.06	0.07	0.08	0.06
Loan	0.12	0.12	0.09	0.10	0.12
Demo	0.04	0.05	0.05	0.03	0.06
Residual	**0.50**	**0.54**	**0.54**	**0.53**	**0.57**

Note: I regress the observed transaction rates on a set of fixed effects and discretized continuous control variables: $r_i = \alpha + \sum_{x=1}^{4} \sum_{j=1}^{J_x} \beta_{xj} \mathcal{I}\{x_i = j\} + \varepsilon_i$, where x refers to lender, state, loan characteristics (LTV and loan amount) and demographic variables (income and age). Group Std. Dev. refers to the standard deviation of β_{xj}'s across categories j. Variance decomposition refers to the share of standard deviation attributed to each group.
Source: Author's calculation using HMDA data.

and regulatory policies.

4.2 Price Discrimination v.s. Risk Adjustments

The observed conditional price dispersion could be explained by two channels. Lenders can use market power to charge borrowers up to willingness to pay. On the other hand, they can adjust the price because they learn about borrowers' private risks. I assume that market structure is orthogonal to the distribution of unobserved risks across borrowers. Using market-level variation in market structure, I show that the market power channel exists and risk adjustment is not the only factor to explain price dispersion.

I estimate the following regression at the market and year level

$$\log Y_{mt} = \beta \log HHI_{mt} + X_{mt}\gamma + \lambda_m + \lambda_t + \varepsilon_{mt}$$

where market m is a combination of the geographical market (state) and demographic market (borrowers' income-quintile at the national level), HHI_{mt} is the Herfindahl–Hirschman index in market m and year t, X_{mt} includes the mean loan amount, mean loan-to-ratio (LTV) in market m and year t, λ_m and λ_t stand for the market and year fixed effects respectively, Y_m refers to the log average interest rate, log standard deviation of interest rate, and log ratio of std. dev. and the mean rate (Coefficient of Variation) in the market m and year t.

Table 3: Price Dispersion and Market Concentration

	Log Interest Rate	Log Std. Dev.	Log Coef of Variation
Log HHI	-0.001 (0.002)	0.016 (0.006)	0.018 (0.005)
R^2	0.91	0.31	0.45

Table 3 shows estimates of the coefficient β from regression 4.2 using HMDA data. Columns (2) and (3) of the table show a positive association between market concentration and price dispersion conditional on market, borrower, and loan characteristics measured both in standard deviation and coefficient of variation. The results thus suggest that market structure is a potential channel to explain conditional price dispersion besides unobserved risks.

4.3 Adverse Selection

This section presents empirical evidence consistent with adverse selection in the mortgage market. Adverse selection implies that borrowers who are willing to accept higher prices are also likely to be unobservably riskier. Ideally, we want to test if borrowers who have a higher willingness to pay (accepting higher prices) are also unobservably riskier (more likely to default). Since I do not observe borrowers' default behavior, I use lenders' loan denial decisions to proxy borrowers' riskiness. If the borrowers are also more likely to be rejected by lenders, it suggests that lenders are cautious about extending credit to individuals they view as risky [7]. An additional challenge is that I do not observe interest rates for loan applications that are denied. To address this challenge, I exploit a discontinuity in interest rates around the loan-to-value (LTV) cutoffs (Tsai, 2023) and use this discontinuity in interest rates to study application denials at the cutoffs. It has been documented that mortgage prices are impacted by the GSE's price-setting rules which follow a step function in LTV and credit scores (Agarwal et al., 2018). The pricing rule hence generates

[7] This could be due to concerns about the borrower's ability to repay the loan, based on indicators of their financial health.

a quasi-random variation in interest rates for borrowers around the cutoffs. I find that both interest rates and rejection rates jump up at the cutoffs, consistent with adverse selection in the market.

Discontinuity in Interest Rates I examine discontinuity in interest rates at potential GSE LTV cutoffs rounded to the nearest 10. I regress interest rates on relative LTV to cutoff dummies and other loan characteristics:

$$r_{imt} = \sum_{k=-2}^{k=2} \beta_k \times \mathcal{I}\{LTV_i = c_i + k\} + x_i'\beta_x + c_i \times \lambda_{mt} + \varepsilon_{imt} \tag{1}$$

where r_{imt} is the interest rate of borrower i's loan in market m and time t, LTV is loan-to-value ratio, c_i are LTV cutoffs. Next, x_i contains borrower/loan characteristics including borrowers' income, loan purpose (purchase or refinance), and property type (single-family site built, manufactured home, etc.). Moreover, I include market-year interactions with cutoff fixed effects to control for variation in absolute value differences in interest rates across cutoffs. The coefficients of $\mathcal{I}\{LTV_i = c_i + k\}$ capture the average interest rates around cutoffs relative to $k = -1$, i.e. normalizing β_{-1} to 0.

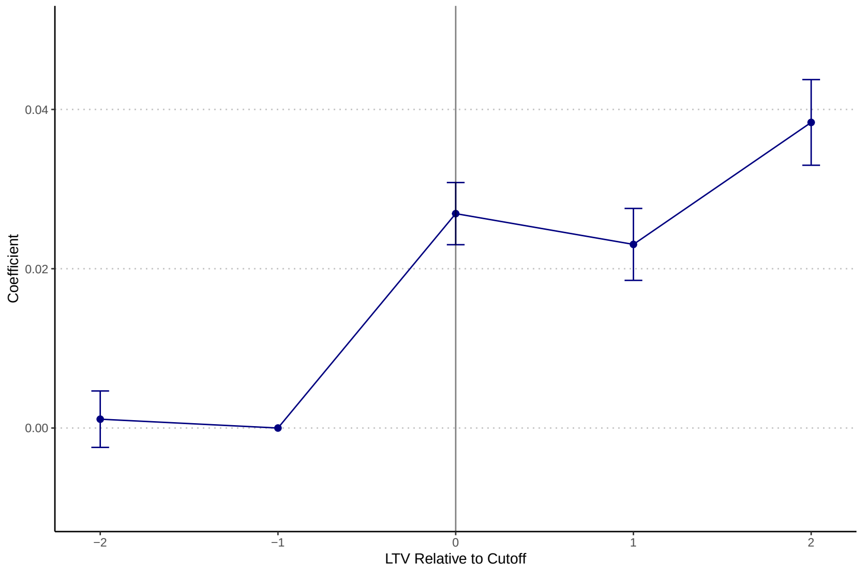
Figure 2 shows the coefficients of $k = -1$ from equation 1. There is a discontinuous jump in interest rates around the cutoffs. This discontinuity is potentially explained by the jump in GSE's guarantee fees which affects the costs of most conforming loans which is exogenous to lenders' and borrowers' decisions. Thus, this discontinuity generates quasi-random variation in interest rates that I exploit to test for adverse selection in the market.

Discontinuity and Loan Denial Rate Using the quasi-random variations in interest rates generated by the GSE pricing rule, I estimate the following RD specification:

$$Y_{imt} = \gamma_D D_i + \gamma_C(LTV_i - c_i) + \gamma_X D_i \times (LTV_i - c_i) + x_i'\beta_x + c_i \times \lambda_{mt} + \varepsilon_{imt} \tag{2}$$

where Y_{imt} is the outcome variable, D_i is an indicator of whether the loan-to-value (LTV) ratio is above a certain cutoff. The coefficient of D_i measures the average differences in outcomes for borrowers who are just above a cutoff and just below a cutoff. I focus on loans that are within ± 2 around the cutoff. My identification assumption is that borrowers' loans that are around a cutoff are

Figure 2: Discontinuity in Interest Rates around Loan-to-Value cutoff



likely to be very similar in both observables and unobservables.

The data documents specific reasons for application denial including debt-to-income (DTI), credit history, etc. I restrict the data to contain only applications that are denied without a specified reason because I want to test for the selection of borrowers based on unobservable characteristics. Applications that are rejected due to DTI and credit history reasons would contaminate the results.

Table 4 shows the results of the regression discontinuity regressions. Column (1) presents results for application denials. The RD coefficient is positive and significant. Given the mean denial rate in the sample is 2%, this 0.002 percentage point increase translates to a 10% increase in application denial rates for borrowers just above the LTV cutoff. The empirical results are consistent with adverse selection: as mortgage rates increase, the marginal borrowers entering the market are associated with higher unobserved risks, implied by higher loan denial rates. Column (2) confirms that there is a jump in interest rates at the GSE pricing cutoffs, generating quasi-experimental variation in interest rates to study adverse selection. Column (3) presents the RD results for log income. It is reassuring that the RD estimate for log income is positive implying borrowers' income is discontinuously higher just above the cutoff. As income is negatively related to both application denial rates and interest rates, this discrete jump will not contaminate the main results. If anything, the main

Table 4: Regression Discontinuity Estimates

	App Denied (1)	Interest Rate (2)	Log Income (3)
RD estimate	0.0023***	0.0285***	0.0397***
	(0.0008)	(0.0057)	(0.0040)
Rel LTV	-0.0019***	-0.0008	0.0154***
	(0.0005)	(0.0035)	(0.0023)
RD $\times$ Rel LTV	0.0039***	0.0036	-0.0676***
	(0.0006)	(0.0049)	(0.0034)
Log Income	-0.0058***	-0.0942***	
	(0.0003)	(0.0041)	
State $\times$ Year $\times$ Cutoff FE	Yes	Yes	Yes
Loan Characteristics	Yes	Yes	Yes
Observations	1,313,370	1,280,760	1,561,266
R^2	0.01792	0.44347	0.12484

Note: Clustered standard errors are in the bracket. Significance levels: ***: 0.01, **: 0.05, *: 0.1.

results would be even greater if the borrowers below and above the cutoff had similar incomes. In sum, I find that borrowers entering with potentially higher interest rates are more likely to be rejected due to lenders' uncertainty about their risks.

4.4 The Need for a Model

Results in this section provide suggestive evidence of conditional price dispersion and adverse selection in the mortgage market. Quantifying the welfare effects of limiting lenders' ability to set individualized prices and varying the extent of information asymmetry is challenging because the market structure is endogenously determined by both demand and supply. Without meaningful exogenous regulatory changes, perhaps the best way to study the effects of different regulations on how lenders set interest rates is to build a structural model of the mortgage market.

In the next section, I present and estimate a structural model of the US mortgage market that features price discrimination and adverse selection discussed in this section. The model will allow me to quantify the impact of the demand and cost channels that explain conditional price dispersion and simulate counterfactual policies.

5 Model Exposition

In this section, I develop a static structural model that features: (i) borrowers' demand for mortgages, (ii) the price setting of mortgages by lenders based on observed signals, and (iii) a flexible correlation between signals, demand, and cost shocks.

5.1 Markets and Players

There are M markets and T time periods, indexed by m and t. I define a market as a combination of demographic characteristics (borrowers' income quintile at the national level) and geographic areas (state), yielding a total of 255 markets. A time period consists of a year, with a total of five years. In each market m and period t, there are N_{mt} heterogeneous households (thereafter borrowers) indexed by i searching for mortgages for the purchase of a residential property. Each borrower can only be in one market in the studied period. There is no repeated borrowing, i.e. I exclude refinances in this study. Lenders are banks or independent mortgage lenders indexed by $j = 1, ..., J_{imt}$. All loans in the study are 30-year fixed-rate loans.

5.2 Timing

Borrowers determine the amount to borrow before searching for mortgages. On the supply side, observe signals about borrowers' private risk types and set individualized prices to optimize expected profit over unobserved demand and cost shocks. A key feature of this model is that the price signals and demand and cost shocks are jointly correlated, so borrowers with higher private risks may have higher costs on borrowers (adverse selection). On the demand side, borrowers choose lenders that maximize their indirect utility among those available in their borrower-specific choice sets.

5.3 Demand: Borrower Discrete Choice Problem

The indirect utility of such a potential borrower i with lender j in market m is given by

$$U_{ijmt} = \alpha r_{ijmt} + \delta_{jmt} + Y_i'\eta + \xi_i^D + \nu_{ijmt} \tag{3}$$

where r_{ijmt} is the mortgage rate offered by lender j to borrower i in market m and year t, Y_i represents borrower level determinants of demand, and ξ_i^D represents borrowers' private information unobserved by the lenders. I describe the information structure about ξ_i^D more in detail in Section 5.5 and 6.3. The lender-market-year fixed effects are given by

$$\delta_{jmt} = \beta_0 + X'_{jmt}\beta + \zeta_{jmt} \tag{4}$$

where X_{jmt} include lender-market-year level determinants of demand and ζ_{jmt} refers to unobserved lender-market-year characteristics.

Borrower i chooses a lender j from his choice set J_{imt} to maximize utility, or else they choose to borrow from the outside option [8].

$$s_{ijmt} = Pr(U_{ijmt} \geq U_{ij'mt}, \forall\ j' \in J_{imt}) \tag{5}$$

Choice Sets The choice set is defined by lenders that originated mortgage loans to borrowers in the same loan type (conforming vs. jumbo), income quintile, age group, gender, state, and year.

5.4 Supply: Lender Mortgage Pricing

Each market m and year t contains J_{mt} lenders that are profit-maximizing firms selling differentiated mortgages to borrowers. They maximize expected profits by setting individualized interest rates for mortgage loans.

Demand signals Lenders observe demand signals ω_i^P and use them to infer borrowers' unobserved types. I define ω_i^P as the residualized interest rate from the following price prediction equation:

$$r_{ijmt} = \lambda_{jmt} + \pi^P \tilde{Y}_i + \omega_i^P \tag{6}$$

[8] I define outside option to be small lenders that have market share below 5%.

where λ_{jmt} are lender-market-year fixed effects, $\tilde{Y}_i$ are observable borrower covariates that include loan-to-value (LTV) ratio, and ω_i^P represents the borrower-specific demand signals [9].

Cost shocks I use cost shocks γ_i^C to represent credit risks that can affect marginal costs but are not observed by lenders. I assume that lenders' marginal costs take the following form:

$$c_{ijmt} = X_j' \beta^C + \lambda_{mt} + \pi^C \tilde{Y}_i + \gamma_i^C \tag{7}$$

where X_j are observable lender covariates that include whether the lender is a nonbank, $\tilde{Y}_i$ are observable borrower covariates that include loan-to-value (LTV) ratio and borrower's log income re-centered around market average, λ_{mt} are market-year[10] fixed effects respectively, and γ_i^C are borrower-specific cost shocks. The market and year fixed effects are important for the identification of cost channels under the assumption that the market power channel is constant within a market and a year and only varies across markets whereas unobserved costs do not vary systematically across markets.

Price Setting Lenders compete à la Nash-Bertrand by setting individualized interest rates that maximize the expected profit over γ_i^C and ξ_i^D. The loan amount is normalized to one and is not affected by changes in interest rates. Lender j's expected profit of originating a product j to borrower i in market m and year t is

$$\Pi_{ijmt}(\omega_i) = \mathop{\mathbb{E}}_{\xi,\gamma}[s_{ijmt}(\xi) \times (r_{ijmt} - c_{ijmt}(\gamma)) \mid \omega_i] \tag{8}$$

where s_{ijmt} is borrower i's probability of choosing product j in market m; r_{ijmt} is mortgage rates; c_{ijmt} denotes marginal cost.

Lenders' optimal pricing equation is given by the first-order conditions

$$\int \frac{\partial s_{ijmt}}{\partial r_{ijmt}} \left(r_{ijmt}(\omega_i) - c_{ijmt}(\gamma)\right) + s_{ijmt}(\xi)\; dF(\xi, \gamma | \, \omega_i) = 0 \tag{9}$$

[9] I cannot observe the same borrower borrowing multiple loans in the data. Therefore, I assume that each borrower only makes one borrowing. The borrower i's level is thus equivalent to the loan level.

[10] With a slight abuse of notation, in lenders' marginal cost function, a market refers to only a geographic market.

Assuming each lender offers *one* representative mortgage product to each borrower, the FOC is reduced to

$$r_{ijmt}(\omega_i) = \underbrace{\frac{\int c_{ijmt}(\gamma) s_{ijmt}(\xi)(1 - s_{ijmt}(\xi))\ dF(\xi, \gamma|\omega_i)}{\int s_{ijmt}(\xi)(1 - s_{ijmt}(\xi))\ dF(\xi|\omega_i)}}_{\text{Marginal Cost}} + \underbrace{\frac{\int s_{ijmt}(\xi)\ dF(\xi|\omega_i)}{-\alpha_q \int s_{ijmt}(\xi)(1 - s_{ijmt}(\xi))\ dF(\xi|\omega_i)}}_{\text{Markup}} \tag{10}$$

The optimal rate is a function of price signals ω_i and depends on the marginal costs and mark-ups.

5.5 Information Structure and Adverse Selection

The model distinguishes between information observed by both lenders and borrowers and information private to each borrower. The former includes hard information in the form of borrower and loan characteristics, as well as, soft information known to the lenders through interaction with the borrowers proxied by ω_i^P. Private information known to the borrowers but not to the lenders is captured by unobservables ξ_i^D and γ_i^C. I assume that ω_i^P, ξ_i^D, and γ_i^C are fixed borrower attributes that don't vary across lenders. This is an intuitive assumption as it implies that risky borrowers have a high demand for mortgages from all lenders, and not differently across different lenders. The signals and shocks are distributed according to the following multivariate normal distribution

$$\begin{pmatrix} \xi_i^D \\ \gamma_i^C \\ \omega_i^P \end{pmatrix} \sim \mathcal{N}\left(\begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \sigma_\xi^2 & \rho_1 \sigma_\xi \sigma_\gamma & \rho_2 \sigma_\xi \sigma_\omega \\ \rho_1 \sigma_\xi \sigma_\gamma & \sigma_\gamma^2 & \rho_3 \sigma_\gamma \sigma_\omega \\ \rho_2 \sigma_\xi \sigma_\omega & \rho_3 \sigma_\gamma \sigma_\omega & \sigma_\omega^2 \end{bmatrix} \right) \tag{11}$$

The joint distribution of the demand shocks ξ_i^D and cost shocks γ_i^C conditional on signals ω_i^P observed by lenders is given by

$$\begin{pmatrix} \xi_i^D \\ \gamma_i^C \end{pmatrix} \mid \omega_i^P \sim \mathcal{N}\left(\begin{bmatrix} \rho_2 \frac{\sigma_\xi}{\sigma_\omega} \\ \rho_3 \frac{\sigma_\gamma}{\sigma_\omega} \end{bmatrix} \omega_i, \begin{bmatrix} (1 - \rho_2^2)\, \sigma_\xi^2 & (\rho_1 - \rho_2 \rho_3)\, \sigma_\xi \sigma_\gamma \\ (\rho_1 - \rho_2 \rho_3)\, \sigma_\xi \sigma_\gamma & (1 - \rho_3^2)\, \sigma_\gamma^2 \end{bmatrix} \right) \tag{12}$$

I interpret the positive correlation ρ_1 between the borrower-specific unobservables driving demand and lenders' costs as evidence of *unconditional* adverse selection: a positive correlation between ξ_i^D and γ_i^C implies that borrowers with a higher unobservable propensity to demand mortgages are also more likely to have higher costs to lenders. Furthermore $\rho = \rho_1 - \rho_2\rho_3$ measures the extent of adverse selection *conditional* on lenders' soft information ω_i^P.

My model is similar to Crawford et al. (2018) but differs in the specification of correlation structure and the assumptions of price signals. First, a novel component of my model is that it does not require loan performance data. Instead, I model adverse selection as a correlation between unobserved demand and cost shocks. One limitation of this setup is that it does not allow for the presence of a moral hazard [11] which refers to borrowers exerting less effort under high repayment requirements. Second, my model does not require panel data with repeated observations for the same borrower. To overcome this data limitation, I assume that the proxies for price signals are random effects instead of fixed effects and I estimate the proxies as the residuals of a price prediction equation.

6 Estimation and Identification

In estimating the structural model described above, it is crucial to consider the different information sets of the borrowers, lenders, and us as researchers. I assume that there are three sets of information: (1) information known by all borrowers, lenders, and us as researchers (*public information*); (2) factors known by borrowers and lenders, but not us as researchers (*soft information*); as well as (3) factors known only by borrowers, but not lenders or us as researchers (*private information*). Distinguishing (2) and (3) is of particular importance. I describe how I do so in what follows.

I deal with two main identification challenges. First, price is an endogenous variable jointly determined by demand and supply. I use a cost shifter as an instrument to identify the price coefficient for borrowers' demand. Second, the identification of adverse selection - the correlation between private demand and cost shocks - requires careful controls of lenders' soft information (those known to lenders but not by us as researchers). For example, a borrower characteristic

[11] Adams et al. (2009) discuss a method to study moral hazard in the lending market using loan performance data.

that is seen by lenders and therefore used for pricing also impacts borrowers' demand. We could obtain a correlation between private demand and cost shocks not due to adverse selection that I seek to capture. I control for this soft information using ω_i^P described in the above section. I discuss this more in detail in Section 6.3.

6.1 Overview

There are three sets of parameters to be estimated: (1) demand parameters $\{\tilde{\delta}, \eta, \alpha, \beta, \pi\}$; (2) correlation parameters $\{\rho_1, \rho_2, \rho_3\}$; and (3) variance parameters $\{\sigma_\omega, \sigma_\xi, \sigma_\gamma\}$.

I estimate the structural parameters in the following three steps:

- **Price Prediction** Use ordinary least squares (OLS) with the price prediction Eq. 6 to estimate ω_i as residuals and obtain predicted prices
- **Step 1** Use a simulated maximum likelihood estimator (SMLE) with the borrowers' choices Eq. 17 to recover $\theta_1 = \{\tilde{\delta}, \sigma_\xi, \rho_2\}$
- **Step 2** Use an IV regression Eq. 19 to estimate the price coefficient and recover the remaining demand parameters $\theta_2 = \{\alpha, \beta, \pi, \eta\}$
- **Step 3** Use non-linear least squares (NLS) estimator with lenders' optimal pricing Eq. 10 to recover $\theta_3 = \{\sigma_\gamma, \rho_1, \rho_3\}$

6.2 Price Prediction

A key challenge I face when estimating the demand model is that I only observe transaction rates if the borrower chooses to borrow. Therefore, I need to predict prices that each borrower faces at all other lenders in their choice sets.

Using the price prediction model in Eq. 6, I can compute predicted interest rates charged to borrower i by lender j in market m as follows:

$$\tilde{r}_{ijmt} = \tilde{r}_{jmt} + \tilde{\pi}^P \tilde{Y}_i + \tilde{\omega}_i^P \tag{13}$$

where $\tilde{r}_{jmt} = \tilde{\lambda}_{jmt}$ is the lender-market-year specific component of the predicted price, $\tilde{Y}_i$ are observable borrower-level covariates, and $\tilde{\omega}_i^P$ is the estimated residuals. With combinations of the estimated fixed effects $\tilde{r}_{jmt}$, coefficients $\tilde{\pi}_i^P$ and residuals $\tilde{\omega}^P$, I can predict the interest rates $\tilde{r}_{ijmt}$ borrower i by lenders they could have chosen but did not.

The OLS results are shown in Table B.1. I confirm that interest rates decrease monotonically with loan amounts. The fit of the regression, as measured by the R^2, increases marginally to 0.57 as I move from separate lender, state, and year fixed effects to interactions of the three variables.

6.3 First Step

I assume that any determinants of demand that are observed by lenders but unobserved by us as econometricians will be accounted for when setting interest rates. I also assume that such information can be summarized by a variable ω_i^D (unobserved to us). I cannot estimate ω_i^D directly, however, I can use borrowers' demand signal $\tilde{\omega}_i^P$(residualized price) estimated in the price prediction equation 6 as a proxy for this demand unobservable. I relate this soft information impacting demand and soft information impacting price through the following equation:

$$\omega_i^D = \eta_2 \tilde{\omega}_i^P \tag{14}$$

Using this relationship, I define all borrower-level covariates influencing demand as

$$\begin{aligned} Y_i \eta &= \eta_1 \tilde{Y}_i + \omega_i^D \\ &= \eta_1 \tilde{Y}_i + \eta_2 \tilde{\omega}_i^P \end{aligned} \tag{15}$$

where $\tilde{Y}_i$ are observable borrower covariates and the second line imposes my assumption on the relationship between demand and pricing unobservables.

Substituting the price prediction equation into the demand equation, we obtain the following demand utility

$$\begin{aligned} U_{ijmt} &= \alpha r_{ijmt} + \delta_{jmt} + Y_i \eta + \xi_i^D + \nu_{ijmt} \\ &= \alpha \left(\tilde{r}_{jmt} + \tilde{\pi}^P \tilde{Y}_i + \tilde{\omega}_i^P \right) + \delta_{jmt} + \left(\eta_1 \tilde{Y}_i + \eta_2 \tilde{\omega}_i^P \right) + \xi_i^D + \nu_{ijmt} \\ &= \underbrace{(\alpha \tilde{r}_{jmt} + \delta_{jmt})}_{\tilde{\delta}_{jmt}} + \underbrace{(\alpha \tilde{\pi}^P + \eta_1)}_{\tilde{\eta}_1} \tilde{Y}_i + \underbrace{(\alpha + \eta_2)}_{\tilde{\eta}_2} \tilde{\omega}_i^P + \xi_i^D + \nu_{ijmt} \\ &= \tilde{\delta}_{jmt} + V_i + \nu_{ijmt} \end{aligned} \tag{16}$$

where $\tilde{\delta}_{jmt} = \alpha \tilde{r}_{jmt} + \delta_{jmt}$ is lender-market-year level fixed effects, $V_i =$

$\eta_1 \tilde{Y}_i + \eta_2 \tilde{\omega}_i^P + \xi_i^D$ captures borrower-level observable covariates, soft information and private information, and ν_{ijmt} is distributed as a Type I Extreme Value. Because the price parameter α does not enter independently in Eq. 16, I cannot identify it in the first step estimation. Instead, I estimate it using an IV method in Step 3 that uses variation in average interest rates at the lender-market-year level $\tilde{r}_{jmt}$.

Based on these assumption, the probability that consumer i borrowers from lender j in market m and year t is

$$s_{ijmt}(\omega_i) = \int_{\xi} \frac{\exp\left(\tilde{\delta}_{jmt} + V_i\right)}{1 + \sum_{k \in J_{imt}} \exp\left(\tilde{\delta}_{kmt} + V_i\right)} \, dF(\xi|\omega_i) \tag{17}$$

The log-likelihood of borrower i choosing lender j in market m is given by

$$\log \mathcal{L}_i = \sum_{j=0}^{J_{imt}} d_{ijmt} \log(s_{ijmt}(\omega_i)) \tag{18}$$

where d_{ijmt} is a dummy for whether borrower i chooses lender j in market m and year t, and $s_{ijmt}(\omega_i)$ represents borrowers' choice probabilities defined in Eq. 17.

Because the number of fixed effects is large in my model, the canonical MLE method is computationally intensive. Borrowing from Chen et al. (2021), I use a Maximization-Minorization (MM) algorithm to proxy the MLE results. I extend Chen et al. (2021) by allowing for random coefficients. In Section A.1, I derive a transfer minorization for the log-likelihood with random coefficients and show that the MM estimator is consistent under this setup. Next, in Section A.2, I discuss the procedure to estimate $\theta_1 = \{\tilde{\delta}, \sigma_{\xi}, \rho_2\}$.

6.4 Second Step

I use instrumental variables estimation to recover the remaining structural parameters in the demand equation, including the price coefficient α. In step 1, we obtained estimates for the lender-market-year level fixed effects $\hat{\tilde{\delta}}_{jmt}$. These constants contain the lender-market-year level covariates X_{jmt}, as well as the lender-market-year components of the predicted prices $\tilde{r}_{jmt}$. I use the estimated constants $\hat{\tilde{\delta}}_{jmt}$ as dependent variables in a linear IV regression on

X_{jmt} and $\tilde{r}_{jmt}$, using cost shifters as instruments:

$$\hat{\hat{\delta}}_{jmt} = \alpha_0 + \alpha_q \tilde{r}_{jmt} + \beta_1 NB_j + \beta_2 X'_{jmt} + \zeta_{jmt} \tag{19}$$

where NB_j denotes non-banks, X_{jmt} are lender-market-year observable covariates, and ζ_{jmt} captures unobserved lender-market-year characteristics. I allow the price coefficient α_q to be heterogeneous across borrowers' income quintile.

Identification Price endogeneity may come from ζ_{jmt} that can include borrowers' unobserved valuation of a lender's brand and quality. To address the issues of price endogeneity, I use the average interest rates set by the lender in other geographical markets in the same year as an instrument for interest rates (Nevo (2001), Hausman and Taylor (1981)).

6.5 Third Step

In Step 1, I described procedures to estimate the variances of private demand information ξ_i^D and the correlation between price signals ω_i^P and ξ_i^D. The remaining distributional parameters to estimate include the variance of private cost shocks γ_i^C and the correlation between price signals ω_i^P and γ_i^C and the correlation between γ_i^C and ξ_i^D. Identification of these parameters comes from lenders' optional pricing equation, specifically, how demand shocks ξ_i^D and γ_i^C jointly influence lenders' pricing.

With the information structure assumption described in Section 5.5 and lenders' optimal pricing equation Eq. 10 , I recover the distributional parameters $\theta_3 = \{\sigma_\gamma, \rho_1, \rho_3\}$. I minimize the distance between the model predicted price using Eq. 10 and the observed transaction rates:

$$\min_{\theta_3} h\left(d_{ijmt} \cdot r_{ijmt}^{model}(\theta_3) - r_{ijmt}^{data}\right)' h\left(d_{ijmt} \cdot r_{ijmt}^{model}(\theta_3) - r_{ijmt}^{data}\right) \tag{20}$$

where d_{ijmt} is a dummy for whether borrower i chooses lender j in market m and year t and $h(.,.)$ is a distance function.

6.6 Identification

In this subsection, I briefly explain the model features and the data variation that identify the model parameters. Variance in observed interest rates helps identify ω_i, parameters related to lenders' signal which I assume to be residuals from a linear price prediction equation. Borrowers' choices across lenders help identify the variance in unobserved determinants of demand σ_ξ. Borrowers internalize that lenders' signal ω_i affects pricing and hence the correlation between unobserved demand shock ξ and signal ω help identify the effectiveness of lenders' signal $\rho_{\xi\omega}$. Moving to the parameters related to unobserved cost determinants. I first use the residuals from the regression of marginal cost c_i on covariates to proxy σ_γ, the standard deviation of γ. With the fixed σ_γ, the correlation between price and signals helps identify the effectiveness that the signals inform about cost $\rho_{\gamma\omega}$ and the change in prices identifies the correlation between ξ and γ, the adverse selection parameter.

7 Estimation Results

In this section, I discuss parameter estimates. I start with demand model parameters and then move lenders' parameters and finally to the covariance matrix.

7.1 Demand Parameters

Table 5 reports estimates of demand parameters by borrowers' income quintile. The average price coefficient is -1.92. The corresponding average demand elasticity is -3.72. High-income borrowers are slightly more elastic to interest rates than low-income. I find mostly positive but not significant coefficients on the *Nonbank* dummy, suggesting that the results are inconclusive about borrowers' preferences for lenders' institutional types conditional on price.

Table 5: Demand Estimates

Income Quintile	1	2	3	4	5 (H)
Interest Rate Elasticities	-1.66	-1.75	-1.90	-2.12	-2.27
	(0.68)	(0.70)	(0.72)	(0.75)	(0.68)

7.2 Supply Parameters

Table 6 reports the marginal costs and mark-ups for banks and nonbanks. On average, banks' marginal costs are 3.2 p.p. and mark-ups are 0.54 p.p., about 16.8%. Nonbanks have slightly lower marginal costs of 3.02 p.p., and the mark-ups are about 17.8%.

Table 6: Estimates for Lenders' Marginal Costs and Mark-ups

	Marginal Costs	Mark-ups
All	3.07	0.54
Bank	3.20	0.54
Nonbank	3.02	0.54

Figure B.1 reports the lenders' marginal costs and mark-ups by borrowers' income quintile. Lenders' marginal costs decrease monotonically by borrowers' income. Nonbanks' marginal costs stay below banks' across all income groups. Mark-ups are relatively constant both across income groups and lender types.

7.3 Covariance Matrix

Table 7 presents the estimates of the covariance matrix. The estimated standard deviation of the demand shock ξ_i and the cost shock γ_i are 0.53 and 0.58 respectively. The estimated $\rho_{\xi\gamma}$ is 0.43 which is consistent with adverse selection: the more costly types are also those that have higher demand for mortgages. Moreover, the correlation parameter conditional on lenders observing soft information ω_i^P, $\rho_{\xi\gamma} - \rho_{\xi\omega}\rho_{\gamma\omega}$, is 0.30. This suggests that the ability to price based on soft information about the borrowers alleviates the extent of adverse selection in the mortgage market. The estimates of $\rho_{\xi\omega}$ and $\rho_{\gamma\omega}$ represent the effectiveness of lenders' signals on learning about demand and cost unobservables ξ_i and γ_i. They are estimated to be 0.13 and 0.91 respectively. This suggests that lenders' signals that they use to set prices are more effective for hidden cost information and less so for borrowers' demand. Thus, the conditional price dispersion in the mortgage market is more from cost adjustments rather than demand-based price discrimination.

8 Counterfactual Scenarios

I run counterfactual experiments to quantify the effects of individualized pricing using soft information and examine its relationship with adverse selection. In the first counterfactual, I analyze the impact of limiting lenders' ability to price using soft information ω_i^P on prices, demand, consumer welfare, and

Table 7: Covariance Matrix Estimates

Covariance Matrix	Demand (ξ)	Cost (γ)	Signal (ω)
Demand (ξ)	$\sigma_\xi = 0.53$		
Cost (γ)	$\rho_{\xi\gamma} = 0.41$	$\sigma_\gamma = 0.58$	
Signal (ω)	$\rho_{\xi\omega} = 0.13$	$\rho_{\gamma\omega} = 0.91$	$\sigma_\omega = 0.48$
Adverse Selection	Unconditional	Conditional	
	$\rho_{\xi\gamma} = 0.41$	$\rho_{\xi\gamma} - \rho_{\xi\omega}\rho_{\gamma\omega} = 0.30$	

Note: $\rho_1 = \rho_{\xi\gamma}$, $\rho_2 = \rho_{\xi\omega}$, and $\rho_3 = \rho_{\gamma\omega}$.

lenders' profits. I examine how the outcomes vary as I restrict lenders to only price based on observable characteristics.

In the second counterfactual simulation, I investigate the effects of an increase in adverse selection on equilibrium prices, demand, and in particular on the dispersion of interest rates. There have been studies that document an increase in price with an increase in adverse selection Crawford et al. (2018). My goal is to further the literature by analyzing how its impact is distributed across borrowers.

8.1 The Effects of Limited Pricing on Soft Information

In the first counterfactual scenario, lenders can no longer charge different prices to observably similar borrowers using soft information. Under the counterfactual scenario, lenders are restricted to charging the same interest rate to all borrowers/loans with the same characteristics X_i [12]. I denote each combination of X_i as a group g. Hence, prices are constant within g.

Lenders optimize the expected profit by choosing a price for each group g:

$$\max_{r_{gjmt}} \Pi_{gjmt}(X_g) = E_g[s_{ijmt}(\xi) \times (r_{ijmt} - c_{ijmt}(\gamma)) \mid X_g] \tag{21}$$

The FOC under the new profit function is given by:

$$r_{gjmt} = \underbrace{\frac{\int c_{ijmt}(\gamma) s_{ijmt}(\xi)(1 - s_{ijmt}(\xi))\, dF(\xi,\gamma)}{\int s_{ijmt}(\xi)(1 - s_{ijmt}(\xi))\, dF(\xi)}}_{\text{Marginal Cost}} + \underbrace{\frac{\int s_{ijmt}(\xi)\, dF(\xi)}{-\alpha \int s_{ijmt}(\xi)(1 - s_{ijmt}(\xi))\, dF(\xi)}}_{\text{Markup}} \tag{22}$$

[12] X_i includes the loan amount, loan-to-value ratio, borrowers' income, and age group.

I compute the equilibrium prices using the following procedure. I denote t as the iteration index. For each group g, starting from the current price p_t and demand s_t, I compute the new equilibrium price as follows:

1. Use the new optimal pricing FOC of Eq. 22 to compute new price r_{t+1} [13]

2. With r_t, compute s_{t+1} using the choice probability Eq. 17

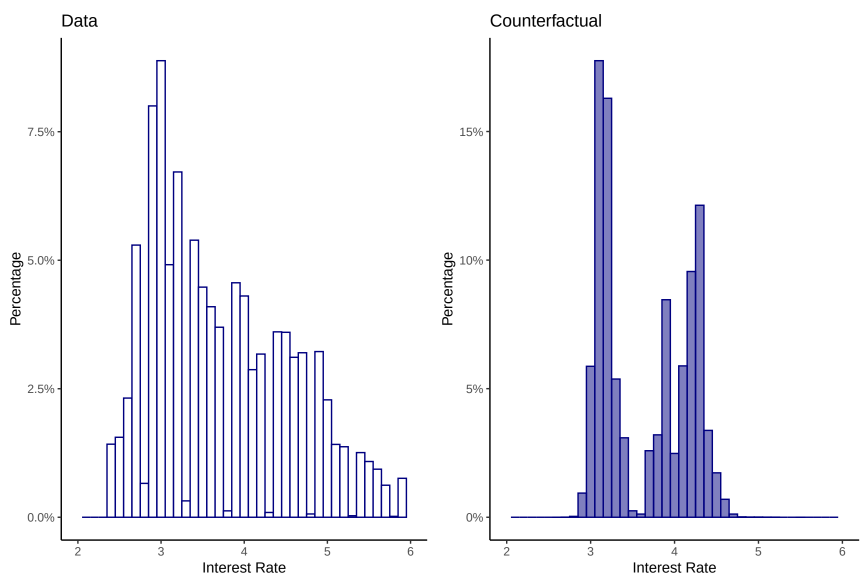
3. Repeat (1) and (2) until convergence

The change in expected consumer surplus is given by:

$$\Delta E[CS_{im}] = \frac{1}{\alpha_q} \ln \left(\sum_{j}^{J} \exp(U^1_{ijmt}) \right) - \ln \left(\sum_{j}^{J} \exp(U^0_{ijmt}) \right)$$

Interest Rates Figure 3 shows the distribution of interest rates in the data and separately in the counterfactual. The distribution of interest rates becomes less individualized. Borrowers who belong to the same group share the same interest rates. The coefficient of variation decreased from 0.21 to 0.17, and the standard deviation decreased from 0.78 to 0.65. These imply a reduction in the dispersion of interest rates.

Figure 3: Changes to the Distribution of Interest Rates

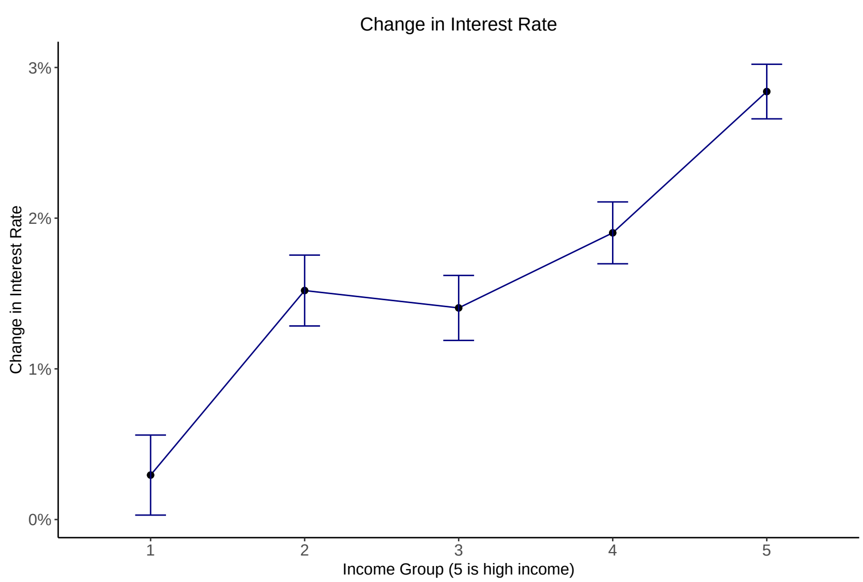


[13]Since the integrals in Eq. 22 does not have a closed form, I simulate the integral using Halton draws.

The sign of change in interest rates is ambiguous and depends on customer groups and market conditions. Individualized pricing allows lenders to charge high prices to high willingness-to-pay and vice versa. Uniform pricing means charging an "average" price to each customer group. Interest rates may decrease as lenders now cannot price discriminate and interest rates may increase as lenders now have to pool consumers into groups. The latter slightly dominates in the counterfactual: average interest rates increased by 1.6%, which translates to approx. $10K increase in total loan cost for a median mortgage at the average rate in the studied period.

The decrease in average interest masks great heterogeneity across customer groups. Figure 4 shows the change in interest rates from baseline to the counterfactual by income group. In the counterfactual, lenders need to pool interest rates across unobserved borrower types. High-income borrowers are generally of low-risk type but are now forced to pool with other risk types. As a result, interest rates increase more among the high-income group.

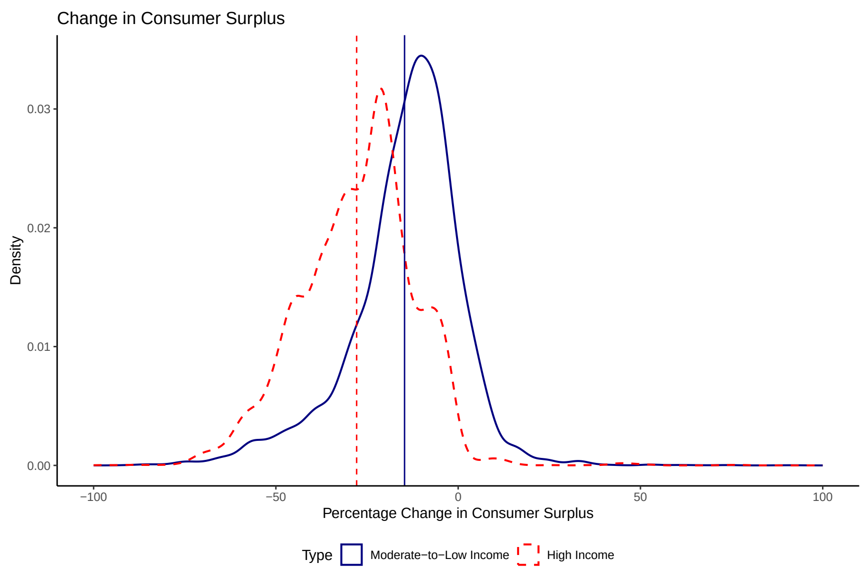
Figure 4: Change in Interest Rates in Baseline and Counterfactual



Consumer Surplus Figure 5 plots the change in consumer plus from the baseline to the counterfactual for the top income (the top 20%) and the moderate-to-low income (the remaining bottom 80%). Relative to the baseline, consumer surplus falls by 18.6% in the counterfactual. This shows that consumers are better off with the status-quo degree of individualized pricing.

As with interest rates, the decrease in consumer surplus masks heterogeneity across borrowers. Consumer surplus falls by 14.7% among the moderate-to-low-income whereas it falls 27.9% among the top-income. In sum, the status quo level of price discrimination particularly benefits high-income consumers.

Figure 5: Distribution of Consumer Surplus in Baseline and Counterfactual



Supply-side Outcomes Lastly, I examine the changes in lenders' outcomes. The big lenders capture even more market share. The market share of the largest lenders (with market shares over the 90^{th} percentile) increased by 1.6 percentage points, implying an increase in market concentration. Lenders' profits increase by 2.6% under the counterfactual compared to the status quo.

9 Concluding Remarks

Price dispersion in the mortgage market has been at the center of policy discussions. Price dispersion is of policy interest, particularly when people who need housing finance the most, often the low-income, cannot obtain it due to market inefficiency. This paper contributes to this debate by providing a quantitative framework to understand the source of price dispersion. In addition, it provides new empirical evidence about how price dispersion is related to information asymmetry, shedding light on potential policy changes that foster information transparency.

To explain this conditional price dispersion, I estimate a structural model that features borrowers' demand for mortgages and lenders' individualized optimal price decisions. I document adverse selection in the form of a positive correlation between the unobserved determinants of borrowers' demand and lenders' costs. Moreover, lenders' ability to use soft information to set individualized prices alleviates the extent of adverse selection. The counterfactual analysis reveals that consumers benefit from the status quo individualized pricing compared to more uniform pricing because it drives down equilibrium prices through enhanced competition. Moreover, I find that the effect is heterogeneous across income groups with higher-income borrowers benefiting more.

Moreover, I study to what extent demand and cost factors contribute to the unexplained price dispersion. This decomposition exercise is of policy interest because adjusting prices based on cost differences mitigates adverse selection and improves efficiency, wherever discriminating consumers based on their demand extracts consumer surplus. If the government's objective is to facilitate housing finance and improve home ownership, particularly for the disadvantaged group, then regulating price discrimination is justified. However, if the dispersion comes from efficient cost adjustments, then regulating is not necessary and can even be harmful. My findings are consistent with cost factors contributing more to lenders' individualized pricing through its relationship with the signals. In sum, my evidence supports light price regulations in the mortgage market as it both allows lenders to adjust risks by learning borrowers' private types and also improves consumer welfare as enhanced competition lowers the equilibrium prices.

Appendices

A MM Algorithm Proofs

I follow Chen et al. (2021) that uses the Minorization-Maximization (MM) algorithm for fast estimation of multinomial logit models with many fixed effects. I extend Chen et al. (2021) to accommodate for random coefficients. With the MM approach, I absorb a large number of fixed effects through linearization and only search over the distributional parameters using maximum likelihood.

A.1 Transfer Minorization

This subsection derives a function S that satisfies the conditions to be a transfer minorization of $\mathcal{L}$.

The log-likelihood function is

$$\mathcal{L}(\theta) = \sum_{j \in J_i} d_{ijm} \log s_{ijm} (\omega_i)$$

where

$$s_{ijmt}(\omega_i) = \int_\xi \frac{\exp\left(\tilde{\delta}_{jmt} + \beta\omega_i + \xi_i\right)}{\sum_{k \in J_i} 1 + \exp\left(\tilde{\delta}_{km} + \beta\omega_i + \xi_i\right)} \, dF(\xi_i|\omega_i)$$

and

$$\xi_i|\omega_i \sim \mathcal{N}\left(\rho_2 \frac{\sigma_\xi}{\sigma_\omega}\omega_i, \ (1-\rho_2^2)\sigma_\xi^2\right).$$

Let S be defined as

$$S\left(\boldsymbol{\theta}; \boldsymbol{\theta}^{(k)}\right) = \mathcal{L}\left(\boldsymbol{\theta}^{(k)}\right) + \frac{1}{2}\sum_j h_j\left(\tilde{\boldsymbol{\psi}}_{im}; \boldsymbol{d}_{im}\right)^2 - \frac{1}{2}\sum_j \left(\tilde{\psi}_{ijm} + h_j\left(\tilde{\boldsymbol{\psi}}_{im}; \boldsymbol{d}_{im}\right) - \psi_{ijm}\right)^2 \tag{23}$$

where

$$h_j(\boldsymbol{\psi}_{im}; \boldsymbol{d}_{im}) = d_{ijm} - \frac{1}{s_{ijmt}} \int_\xi \left(\frac{exp(\psi_{ijm})}{\frac{1}{exp(\xi_i)} + \sum exp(\psi_{ikm})}\right)^2 dF\left(\xi_i \mid \omega_i\right),$$

$$\tilde{\psi}_{ijmt} = \tilde{\delta}_{jmt}^{(k)} + \beta^{(k)}\omega_i,$$

$$\psi_{ijmt} = \tilde{\delta}_{jmt} + \beta\omega_i.$$

Proof: See Appendix A.3.

Eq. (23) is a second-order Taylor expansion of the likelihood function. We search $\boldsymbol{\theta}$ in the iterative MM algorithm, which only appears in the third part of the right-hand side of Eq. (23). Focusing on the third part, we can think of optimization of S as a linear regression of $\tilde{\delta}_{jmt}^{(k)} + \beta\omega_i^{(k)} + h_j(\psi_{ijmt}; d_{ijmt})$ on ω_i and $\tilde{\delta}_{jmt}$. This is the central benefit of the MM approach to canonical MLE as we convert non-linear optimization to sequential linear optimization, which is particularly attractive when there are many regressors. We use fast fixed effects estimation methods that absorb the large number of $\tilde{\delta}_{jmt}$'s.

A.2 Algorithm

The algorithm goes as follows. We begin by guessing the distributional parameters ρ and σ_ξ in the outer loop. Then with the given distributional parameters, we compute the necessary variables to implement the MM algorithm. We use a canonical MLE to search for ρ and σ_ξ in the outer loop.

Outer Loop (MLE) In the outer loop, we recover the distributional parameters $\{\rho_2, \sigma_\xi\}$. We represent them as functions of the demand parameters, i.e. α_m and $\tilde{\delta}_{jmt}$ which we solve for in the inner loop. We estimate $\{\rho_2, \sigma_\xi\}$ using MLE where the log-likelihood is given in Eq. (17).

Inner Loop (MM) Given the distributional parameters $\{\rho_2, \sigma_\xi\}$ from the outer loop, we implement the MM algorithm as follows. We begin by guessing the parameters $\theta^{(0)}$. We iterate on the following sequence of steps, which updates the parameters $\theta^{(k)}$ in iteration k:

1. Minorization Step: Given $\theta^{(k)}$, we compute

$$v_{ijmt}^{(k)} = \tilde{\psi}_{ijt} + h_j\left(\tilde{\boldsymbol{\psi}}_{im}; \boldsymbol{d}_{im}\right).$$

2. Maximization step:

   (a) Update $\beta^{(k)}$ by regressing $v_{ijmt}^{(k)}$ on ω_i with all variables demeaned at the lender-market-year level.

(b) Update $\tilde{\delta}_{jmt}^{(k)}$ by computing

$$\tilde{\delta}_{jmt}^{(k+1)} = \frac{1}{N_{jmt}} \sum_{i \in J \times M} (v_{ijmt}^{(k)} - \beta^{(k+1)} \omega_i) \quad \forall j, m.$$

3. Return to Step 1 for iteration $k+1$ as long as the Euclidean distance between $\boldsymbol{\theta}^{(k+1)}$ and $\boldsymbol{\theta}^{(k)}$ is above some tolerance.

A.3 MM Algorithm Transfer Minorization Proof

$$h_j = \frac{\partial \mathcal{L}}{\partial \psi_{ijm}} = d_{ijm} \frac{1}{s_{ijm}(\omega_i)} \frac{\partial s_{ijm}}{\partial \psi_{ijm}} + \sum_{l \neq j} d_{ilm} \frac{1}{s_{ilm}} \frac{\partial s_{ilm}}{\partial \psi_{ijm}} \tag{24}$$

$$\begin{aligned}
\frac{\partial s_{ijm}}{\partial \psi_{ijm}} &= \int_\xi \frac{\partial \frac{exp(\psi_{ijm}+\xi_i)}{1+\sum exp(\psi_{ikm}+\xi_i)}}{\partial \psi_{ijm}} dF(\xi_i \mid \omega_i) \\
&= \int_\xi \frac{exp(\psi_{ijm}) \left(\frac{1}{exp(\xi_i)} + \sum exp(\psi_{ikm}) \right) - exp(2\psi_{ijm})}{\left(\frac{1}{exp(\xi_i)} + \sum exp(\psi_{ikm}) \right)^2} dF(\xi_i \mid \omega_i) \\
&= s_{ijmt} - \int_\xi \left(\frac{exp(\psi_{ijm})}{\frac{1}{exp(\xi_i)} + \sum exp(\psi_{ikm})} \right)^2 dF(\xi_i \mid \omega_i) \qquad (25)
\end{aligned}$$

For $l \neq j$,

$$\frac{\partial s_{ilm}}{\partial \psi_{ijm}} = \int - \frac{\exp(\psi_{ilm}) \exp(\psi_{ijm})}{\left(\left(\frac{1}{\exp(\xi_i)} \right) + \Sigma \exp(\psi_{ikm}) \right)^2} dF(\xi_i \mid w_i) \tag{26}$$

Plugging Equation (25) and (26) into Equation (24),

$$h_j = d_{ijmt} - \sum_{l \in J_i} \frac{d_{ilm}}{s_{ilm}} \int \frac{\exp(\psi_{ijm}) \exp(\psi_{ilm})}{\left(\exp(-\xi_i) + \sum_k \exp(\psi_{ikm})\right)^2} dF(\xi_i \mid w_i)$$

Define $DEN = \exp(-\xi_i) + \sum_k \exp(\psi_{ikm})$

$$\begin{aligned}
&\sum_{l \in J_i} \frac{d_{ilm}}{s_{ilm}} \int \frac{\exp(\psi_{ijm}) \exp(\psi_{ilm})}{\left(\exp(-\xi_i) + \sum_k \exp(\psi_{ikm})\right)^2} dF(\xi_i \mid w_i) \\
=& \sum_{l \in J_i} \frac{d_{ilm}}{\int \frac{\exp(\psi_{ilm})}{DEN} dF(\xi_i \mid \omega_i)} \int \frac{\exp(\psi_{ijm}) \exp(\psi_{ilm})}{(DEN)^2} dF(\xi_i \mid w_i) \\
=& \sum_{l \in J_i} \frac{d_{ilm}}{\exp(\psi_{ilm}) \int \frac{1}{DEN} dF(\xi_i \mid \omega_i)} \exp(\psi_{ilm}) \int \frac{\exp(\psi_{ijm})}{(DEN)^2} dF(\xi_i \mid \omega_i) \\
=& \sum_{l \in J_i} \frac{d_{ilm}}{\int \frac{1}{DEN} dF(\xi_i \mid \omega_i)} \int \frac{\exp(\psi_{ijm})}{(DEN)^2} dF(\xi_i \mid w_i) \\
=& \frac{1}{\int \frac{1}{DEN} dF(\xi_i \mid \omega_i)} \int \frac{\exp(\psi_{ijm})}{(DEN)^2} dF(\xi_i \mid \omega_i) \sum_{l \in J_i} d_{ilm} \\
=& \frac{1}{\int \frac{\exp(\psi_{ijm})}{DEN} dF(\xi_i \mid \omega_i)} \int \frac{\exp(2\psi_{ijm})}{(DEN)^2} dF(\xi_i \mid \omega_i) \\
=& \frac{1}{s_{ijm}} \int \left(\frac{\exp(\psi_{ijm})}{DEN} \right)^2 dF(\xi_i \mid \omega_i)
\end{aligned}$$

Therefore,

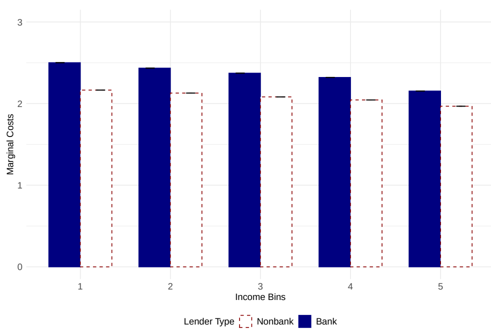
$$h_j = d_{ijm} - \frac{1}{s_{ijmt}} \int_\xi \left(\frac{exp(\psi_{ijm})}{\frac{1}{exp(\xi_i)} + \sum exp(\psi_{ikm})} \right)^2 dF(\xi_i \mid \omega_i)$$

B Supplementary Tables and Figures

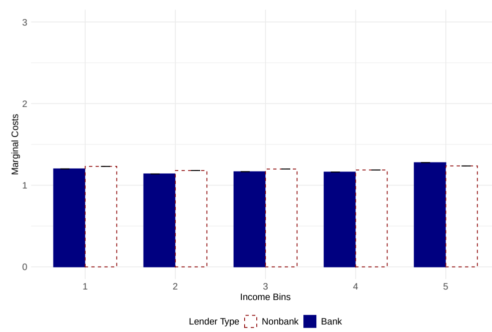
Table B.1: Price regressions

	(1)	(2)	(3)	(4)
Loan amount	-0.0006**	-0.0006**	-0.0005*	-0.0006*
	(0.0002)	(0.0002)	(0.0002)	(0.0002)
Loan-to-value ratio		9.33×10^{-5}	0.0001	0.0001
		(0.0001)	(0.0001)	(0.0001)
Income		4.41×10^{-5}	4.59×10^{-5}	0.0003
		(1.96×10^{-5})	(3.58×10^{-5})	(0.0002)
Age above 62		-0.0193	-0.0385	-0.0394
		(0.0299)	(0.0226)	(0.0208)
Observations	822,078	814,776	814,776	814,776
R^2	0.41628	0.41751	0.53469	0.57069
State fixed effects	✓	✓	✓	
Year fixed effects	✓	✓	✓	
Lender fixed effects			✓	
State-year-lender fixed effects				✓

Figure B.1: Estimates of Marginal Costs and Mark-ups (5 is the highest income)



(a) Marginal Costs



(b) Mark-ups

References

Adams, W., Einav, L., and Levin, J. (2009). Liquidity constraints and imperfect information in subprime lending. American Economic Review, 99(1):49–84.

Agarwal, S., Chomsisengphet, S., Mahoney, N., and Stroebel, J. (2018). Do banks pass through credit expansions to consumers who want to borrow? The Quarterly Journal of Economics, 133(1):129–190.

Alexandrov, A. and Koulayev, S. (2018). No shopping in the us mortgage market: Direct and strategic effects of providing information. Consumer Financial Protection Bureau Office of Research Working Paper, (2017-01).

Allen, J., Clark, R., and Houde, J.-F. (2019). Search frictions and market power in negotiated-price markets. Journal of Political Economy, 127(4):1550–1598.

Banga, H. W. (2022). Price regulation and market power: Evidence from manufactured home loans. Unpublished working paper.

Benetton, M. (2021). Leverage regulation and market structure: A structural model of the uk mortgage market. The Journal of Finance, 76(6):2997–3053.

Bhutta, N., Fuster, A., and Hizmo, A. (2020). Paying too much? price dispersion in the us mortgage market.

Borenstein, S. and Rose, N. L. (1994). Competition and price dispersion in the us airline industry. Journal of Political Economy, 102(4):653–683.

Chen, M., Rysman, M., Wang, S., and Wozniak, K. (2021). Payment instrument choice with scanner data: An mm algorithm for fixed effects in multinomial logit models. Technical report, Working Paper.

Clark, R., Houde, J.-F., and Kastl, J. (2021). The industrial organization of financial markets. In Handbook of Industrial Organization, volume 5, pages 427–520. Elsevier.

Crawford, G. S., Pavanini, N., and Schivardi, F. (2018). Asymmetric information and imperfect competition in lending markets. American Economic Review, 108(7):1659–1701.

Einav, L., Finkelstein, A., and Mahoney, N. (2021). The io of selection markets. In Handbook of Industrial Organization, volume 5, pages 389–426. Elsevier.

Jin, C. (2022). Does competition improve information quality: Evidence from the security analyst market.

Matcham, W. (2022). Risk-based quantity limits in credit card markets. Technical report, Working Paper.

Nelson, S. (2018). Private information and price regulation in the us credit card market. Unpublished working paper.

Robles-Garcia, C. (2019). Competition and incentives in mortgage markets: The role of brokers. Unpublished working paper.

Tsai, H.-T. (2023). Advantageous selection with intermediaries: a study of gse-securitized mortgage loans. The RAND Journal of Economics, 54(4):668–694.