Essays in Industrial Organization with Disaggregate Data

by

Colin Watson

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
(Economics)
in the University of Michigan
2020

Doctoral Committee:

Associate Professor Ying Fan, Chair
Assistant Professor Zach Brown
Professor Francine Lafontaine
Associate Professor Yesim Orhun

Colin Watson

watsonco@umich.edu

ORCID iD: 0000-0003-3059-1393

Acknowledgments

I thank my committee members Profs. Ying Fan, Zach Brown, Francine Lafontaine, and Yesim Orhun for their insightful and patient advising. For valuable conversations in the development of this work, I thank especially Kimberly Conlon, Anirudh Jayanti, Xuan Teng, Travis Triggs, and Andrew Usher. Parts of this work were conducted during an internship at the Consumer Financial Protection Bureau, where I received helpful feedback from Ron Borzekowski, Brian Bucks, and Sergei Koulayev.

TABLE OF CONTENTS

Acknowledgments . . . . . **ii**

LIST OF FIGURES . . . . . **v**

LIST OF TABLES . . . . . **vi**

LIST OF ABBREVIATIONS . . . . . **viii**

Abstract . . . . . **ix**

Chapter

Introduction **1**

1 Price Discrimination Against Inattentive Mortgage Borrowers **4**

1.2 Data 6
1.2.1 Interest rate recall 11
1.2.2 Data on search intensity 15
1.3 Model 18
1.3.1 Simultaneous search implies implausibly weak price preferences . 21
1.3.2 Estimating the sequential search model 22
1.4 Results 26
1.4.1 Interest rate recallers only 26
1.4.2 Full results using both consumer types 27
1.5 Counterfactuals 31
1.6 Features of the mortgage market and robustness 33
1.6.1 Loan performance and FHA insurance 34
1.6.2 Relation to models of rational inattention 38
1.7 Conclusion 39

Appendices **41**

1.A Estimating Price Sensitivity α_m under Simultaneous Search 41
1.B Derivation of the Markup under Rational Inattention 48

2 Consumer Search and Switching Costs in Equilibrium **51**

2.2 Model 53
2.3 More search may increase or decrease prices 56
2.4 Data 57

2.5 Estimation . . . 58
2.5.1 Approximate best response iteration . . . 60
2.6 Results . . . 62
2.6.1 Applicability of the estimates . . . 64
2.7 Comparison with other models of search and switching . . . 66
2.7.1 Search costs . . . 66
2.7.2 Switching costs . . . 67
2.8 Conclusion . . . 69

Appendices . . . 74

2.A Price Derivatives of the Transition Matrix . . . 74

3 Generalized Linear Models for Demand Estimation with Lightly Aggregated Data . . . 76

3.2 Literature . . . 77
3.3 Model and estimation . . . 79
3.3.1 Multinomial logit . . . 80
3.3.2 Extension to mixed logit . . . 81
3.4 Monte Carlo simulations . . . 84
3.5 Conclusion . . . 88

Conclusion . . . 90

Bibliography . . . 91

LIST OF FIGURES

1.1 Density of cost-adjusted interest rate spreads paid, by recall status 12
1.2 Possible probability densities of price offers and firm costs 27
1.3 Density of firm costs 28
1.4 Density of price offers 30
1.5 Earth Mover's Distance as a function of the price sensitivities 30
1.6 Firm's optimal markup as a function of the cost of lending 32
1.A1 Possible probability densities of price offers and firm costs 43
1.A2 Robustness check: Alternative trimming of the cost distribution (2) 46
1.A1 Illustration of the simultaneous search model estimation method, using 8 simulated consumers 47
1.A2 Estimation of the sequential search model with 10 million simulated consumers 48

2.8.1 Firms' total surplus versus the increase in firm price choices when search or switching costs are removed 73

LIST OF TABLES

1.1 Mean cost-adjusted interest rate spread 9
1.2 Summary of the NSMO data 10
1.3 Summary statistics by whether the consumer recalled an interest rate 13
1.4 Share of survey respondents who do not recall the interest rate 15
1.5 Regression of the cost-adjusted rate spread on demographic characteristics of the respondent. Sample: Loans with only one borrower. 16
1.6 Regression of the cost-adjusted rate spread on demographic characteristics of the respondent. Sample: Loans with more than one borrower. 17
1.7 Regression of cost-adjusted rate spread on the number of lenders seriously considered 19
1.8 Demand estimates 29
1.9 Results of counterfactually altering preferences and offers 32
1.10 Regression of delinquency or default on rate recall and the cost-adjusted rate spread 36
1.11 Delinquency rates by FHA versus conventional loans and recall of the interest rate 37
1.12 Robustness check on FHA loans 37
1.A1 Tobit regression to adjust the rate spread for factors affecting the cost or risk of lending 42
1.A2 The 25 most important variables in a random forest model of the cost-adjusted rate spread 44
1.A3 Robustness check: All search intensities doubled 44
1.A4 Robustness check: Counterfactuals with all search intensities doubled 45
1.A5 Robustness check: Alternative trimming of the cost distribution(1) 45
1.A6 Robustness check: Limit to the most creditworthy borrowers 45

2.8.1 Illustration of optimal search sets 70
2.8.2 Results from Model 0 of Honka (2014) 70
2.8.3 Additional parameter estimates from Model 0 of Honka (2014) 70
2.8.4 Summary of cost estimation results under alternative assumptions of how much firms discount future profits 71
2.8.5 Estimated marginal costs $\overline{r}_j$ and actual and projected market shares 71
2.8.6 Expenses of major insurers on claims and underwriting 72
2.8.7 Results of counterfactually varying the search cost c and inertia preference β . 72

3.4.1 Comparison of three estimation methods for a binary choice problem 85

3.4.2 Results of three estimation methods for a multinomial logit choice process . . 87
3.4.3 Results of binomial regression for mixed logit 88

LIST OF ABBREVIATIONS

APOR Average Prime Offer Rate

APR Annual Percentage Rate

BLP Berry, Levinsohn, and Pakes (1995)

CFPB Consumer Financial Protection Bureau

DTI Debt to Income ratio

EMD Earth Mover's Distance

FE Fixed Effects

GLM Generalized Linear Model

GLMM Generalized Linear Mixed Model

HMDA Home Mortgage Disclosure Act

LTV Loan to Value ratio

NSMO National Survey of Mortgage Originations

OLS Ordinary Least Squares

PMMS Freddie Mac Primary Mortgage Market Survey

PTI Payment to Income ratio

ABSTRACT

This dissertation presents several empirical and methodological results in industrial organization, with a focus on settings with microdata or lightly aggregated data. Chapter 1 estimates a model of search and price discrimination in the US home mortgage market, using microdata from two types of consumers to identify the model. Consumers who fail to recall the interest rate pay more for their mortgages, with most of the disparity explained by price discrimination. Chapter 2 estimates an equilibrium model of the US auto insurance market. Consumers face search and switching costs, which firms take into account in their pricing decisions. Counterintuitively, consumers may be harmed in aggregate by lower search costs. Chapter 3 considers the standard problem of estimating logit or mixed logit demand, but in disaggregate data where markets are too small for the market shares to reliably equal the choice probabilities. I adapt binomial regression to estimate a multinomial logit model and show that a version of the Salanié and Wolak (2019) linearization can be applied to binomial regression to approximate the mixed logit model.

Introduction

Industrial organization economists increasingly have access to disaggregate data on the choices and characteristics of individual consumers, or of small groups of consumers. The most apparent benefit of disaggregate data is to address questions about specific groups of consumers who may have different preferences and shopping strategies, and who may in some cases be treated differently by firms. However, research not focusing on consumer heterogeneity may still benefit from disaggregate data as a rich source of identification. Honka and Chintagunta (2017) for instance use microdata to empirically distinguish the sequential and simultaneous search models, a distinction that would not be evident with only aggregate data on prices and quantities. In other cases, disaggregate data is used to answer the same questions as aggregate data but with greater robustness. An example can be seen in two papers on the automobile market: Murry and Zhou (2016) and Moraga-González, Sándor and Wildenbeest (2015). Both papers estimate a travel cost for the consumer to visit and search a dealer (or in Murry and Zhou (2016), an agglomeration of dealers). Murry and Zhou (2016) however take advantage of microdata on dealer and customer locations in their estimation, while Moraga-González et al. (2015) must rely on an assumption that consumers wishing to search vehicles of a certain brand do so at the dealer of that brand nearest to their home.

Disaggregate data also presents challenges for common industrial organization models developed for aggregate data. As discussed in Chapter 3, random utility models following McFadden (1974) and Berry, Levinsohn and Pakes (1995) often rely on the large-market assumption that the observed market shares are equal to the underlying choice probabilities. This assumption may break down when markets are defined at a disaggregate level such as a store or zip code over a few weeks, rather than e.g. the entire U.S. automobile market over a year as modeled by Berry, Levinsohn and Pakes (1995) (hereafter BLP). In other cases, the data may be consumer-level microdata with no imposed aggregation at all, as in Chapters 1 and 2.

This dissertation examines several empirical and (in Chapter 3) methodological questions, with a focus on the use of disaggregate data. The first two chapters apply microdata

(or in Chapter 2, a demand model already estimated from microdata) to estimate equilibrium models of consumer search. Chapter 1 investigates the causes of price dispersion in the US home mortgage market. I show that the interest rate paid by a consumer is strongly associated with whether the consumer was able to recall the interest rate on a survey. Controlling for factors associated with the cost of lending, I estimate a model of sequential search in which firms may price discriminate against inattentive consumers, identified as those who subsequently fail to recall the interest rate. Because inattentive consumers tend to search fewer lenders than their attentive counterparts, price discrimination against inattentive consumers can be rational for firms. With only one type of consumer it would not be possible to identify the strength of interest rate preferences versus preferences for the (unobserved) other characteristics of the loan. However, identification is possible with access to microdata containing (at least) two types of consumers who have different price preferences and/or search costs. I find that most of the interest rate differential paid by inattentive consumers can be explained by price discrimination, with only 10 percent explained directly by consumer behavior.

Chapter 2 studies firms' pricing decisions in light of search and switching costs. Switching costs are generally not identifiable from aggregate data, but can be identified from microdata that reports consumer movements across firms. Using the microdata-derived demand estimates of Honka (2014), I estimate each firm's cost of providing insurance and use the resulting estimates to simulate counterfactuals. While lower search and switching costs would benefit an individual consumer, the counterfactual simulations show that this benefit is eclipsed in aggregate by a supply-side response of increasing prices. Intuitively, low search costs lessen the pressure on firms to compete to be searched. Firms instead compete to be chosen after the consumer has searched them, at which point the firms have market power due to the product differentiation that the consumer's search has revealed. Estimation relies on a combination of methods: automatic differentiation (with respect to price) of a complicated demand function, a Markov chain model of consumer movements across firms, and a linearized version of best response iteration to compute an equilibrium for each counterfactual.

Chapter 3 turns to developing methods for lightly aggregated data, which falls between consumer-level microdata and the large-market datasets traditionally used for demand estimation. I reframe the traditional logit demand estimation problem such that it can be estimated by binomial regression, a fast and standard method in modern statistical software. Compared to aggregate data methods, this approach is (1) robust to small markets in which market shares may not equal choice probabilities and (2) unrestrictive in its fixed effects structure, allowing for specifications that are less likely to overfit in small or disag-

gregate samples. I adapt the Salanié and Wolak (2019) linearization to approximate a mixed logit model at little additional computational cost, while showing that this linearization can be applied without the additional instruments originally required by Salanié and Wolak (2019). Monte Carlo simulations are used to validate binomial regression and compare it to competing methods such as Gandhi et al. (2017).

As observed by Berry and Haile (2020),

> micro data not only permits richer demand specifications but also can substantially soften the reliance on instrumental variables, reducing both the number and types of instruments required.

Yet it is not always apparent how models and methods originally conceived for aggregate data can be adapted to micro or lightly aggregated data, or how to extend an unusually rich microdata-informed demand specification to an equilibrium model. This dissertation aims to provide methods and examples relevant to disaggregate data, and to address questions of search, switching, and price discrimination where microdata shows special promise.

Chapter 1

Price Discrimination Against Inattentive Mortgage Borrowers

Consumers differ in their willingness to spend time and effort searching for the best deal. In a survey of recent mortgage borrowers, half of consumers reported seriously considering only one lender or mortgage broker. Only 15 percent considered three or more. Consumers with higher income and education were more likely to consider additional lenders (Avery et al., 2017). As past work has noted, searching only one or two firms in this high-stakes environment is a puzzling decision that reflects either an extremely high search cost or simple confusion (Woodward and Hall (2012), Alexandrov and Koulayev (2017)).

At a glance, the low level of search might suggest that one mortgage is perceived to be much like another. In reality interest rates are negotiated between borrower and lender, with substantial dispersion in prices. Bhutta et al. (2018) show that price dispersion cannot be fully explained by differences in creditworthiness, discount points paid, or even by the consumer's choice of lender. The 10th to 90th percentile range of interest rates adjusted for these controls remains at 0.48 percentage points. This residual price dispersion may be explained by the mortgage search and negotiation process. In particular, consumers may have different strategies or levels of knowledge in searching for and choosing a lender. To the extent that lenders observe these differences, they may adjust their rate offers or negotiating strategies to price discriminate against consumers who appear more likely to accept a high rate offer.

This paper estimates a search model of the mortgage market using a merged survey and administrative dataset of recent borrowers. This contributes in two ways to the literature on the US mortgage market. First, I show that consumers whose survey responses indicate an inattentive approach to shopping tend to pay much higher interest rates, by as much as 1 percentage point on average. Second, much of this disparity is due to price discrimination

against inattentive consumers. The central assumption underlying the price discrimination result is that consumer inattention is unrelated to the cost of lending after controlling for standard measures of creditworthiness.

The immediate policy implication of price discrimination is that encouraging consumers to search more may be insufficient to improve their market outcomes. Inattentive consumers continue to pay very high rates even when they search more extensively. Consumer education must also empower borrowers to negotiate favorable terms from the lenders they search. The relation between the search and negotiation processes might also be emphasized. For example, the Consumer Financial Protection Bureau's "Owning a Home" guide advises that "Your best bargaining chip is usually having Loan Estimates from other lenders in hand."[1]

The remainder of the paper proceeds as follows. Section 1.2 describes the data and establishes two patterns. First, consumers' ability to recall their interest rate is a strong predictor of paying a low rate and is associated with greater knowledge and attention as shown by other survey responses. Yet interest rate recall is associated with only a slight increase in the number of firms searched, suggesting that the difference in outcomes arises instead from price discrimination or differences in negotiation effectiveness. Second, increased reported search intensity is only weakly associated with paying a lower interest rate.

Section 1.3 presents a structural model of search over price and unobserved nonprice characteristics. I show that the simultaneous search model implies an implausibly weak price preference. This paper therefore uses a sequential search model. In light of the very weak association between search intensity and accepted interest rate, I assume that all consumers (within a type) have the same search cost. Price dispersion in the model is generated by dispersion in lender costs and the idiosyncratic nonprice utility of a firm to different consumers.[2]

With one type of consumer the model is identified only up to the price disutility α. I divide consumers into two types according to whether they recalled the interest rate. I assume that firms may effectively make a different offer to each type, e.g. by leaving an opening for negotiation that inattentive consumers do not exploit. The two consumer types are assumed to differ only in their preferences and search behavior, with any differences in the cost of lending accounted for by controls. These assumptions identify the full model.

Section 1.4 presents results. Relative to the 86 percent of consumers who do recall the

[1] https://www.consumerfinance.gov/owning-a-home/process/compare/fine-tune-loan-offers/

[2] As shown by Diamond (1971), dispersion in firm costs alone would not be enough to generate price dispersion in equilibrium.

interest rate, non-recallers have less than half the price sensitivity and more than twice the search cost in interest rate points. Part of this apparent difference in preferences may be due to confusion among non-recallers about the market environment and a tendency to believe that the rate offers they receive are the best available.

Section 1.5 decomposes the interest rate penalty for non-recallers. Of the recaller/non-recaller difference in mean interest rate spread, 10 percent is a direct effect of preferences and 90 percent is a price discrimination effect. This substantial price discrimination underscores the finding of Bhutta et al. (2018) that there is substantial interest rate dispersion in mortgages within the same lender. Consumer conscientiousness and attention, as proxied by interest rate recall, provide an explanation for why a lender might offer different prices to different consumers.

Section 1.6 discusses some additional features of the mortgage market and their implications for the results. Several robustness checks are presented. Section 1.7 concludes.

1.2 Data

This paper uses the National Survey of Mortgage Originations (NSMO) developed by Avery et al. (2017). NSMO combines a survey of borrower characteristics and search intensity with linked administrative data on lender identity and loan terms. While borrower characteristics and outcomes have been observed separately in past data collection (Woodward and Hall, 2012), the novelty of NSMO is that these variables are now linked in transaction-level microdata, allowing the estimation of models with consumer heterogeneity such as this paper and Alexandrov and Koulayev (2017). Consumer characteristics include self-reported financial and demographic characteristics as well as self-reported familiarity with various parts of the mortgage process. Consumers also report the number of lenders or brokers they chose to "seriously consider." The NSMO data collection is ongoing and the data for this paper spans mortgage originations from 2013 through 2016.

Avery and Borzekowski (2019) introduce a volume of recent research using NSMO. In this volume, Critchfield et al. (2019) find that borrowers in rural areas tend to pay higher rates and to express less confidence about their knowledge of the mortgage process. For example, 55 percent of borrowers in metro areas reported that they had been "very familiar" with the mortgage process when they began this process. Controlling for demographics, borrowers in rural counties were 10 percentage points less likely to express such confidence. I observe that only 83.6 percent of rural borrowers recalled their interest rate versus 86.5 percent of borrowers in a metro area, which may help to explain the higher rates paid by rural borrowers. Other articles in this volume analyze borrowers' expectations of

changes in housing prices (Redmer, 2019), evaluate the impact of a disclosure intervention on shopping behavior (Bucks et al., 2019) and evaluate differences in outcomes between borrowers who took advantage of housing counseling and similar borrowers who did not not (Argento et al., 2019). In a simple comparison of the 908 borrowers who underwent housing counseling to all 14,057 who did not, I find no significant difference in interest rate recall between the groups.

Alexandrov and Koulayev (2017) use a more detailed restricted version of the NSMO data to evaluate the potential gains to consumers of searching more, as well as the effect of brand preferences in causing consumers to accept higher-priced offers. The restricted version of NSMO includes the identity of the lender, which is omitted from the public use version to protect the confidentiality of survey respondents. The model I apply to the public use data necessarily has a much less detailed treatment of non-price preferences. The focus of Alexandrov and Koulayev (2017) is on the choice of how many and which firms to search, while this paper focuses on consumer characteristics and the role of price discrimination in explaining the large dispersion in prices.

Table 1.2 summarizes a few key variables in NSMO.[3] The analysis sample is 30-year fixed rate mortgages, which are 64 percent of all mortgages in the data. The loan term is known from administrative data, but the identification of fixed versus adjustable rate mortgages relies on consumer survey responses. I discard both the 6.8 percent of consumers who reported having an adjustable rate mortgage and the 2.6 percent who did not know whether their rate was adjustable. The share of consumers reporting an adjustable rate mortgage in the data is similar to shares of adjustable rate mortgages reported by the Mortgage Bankers Association during this period.[4]

NSMO's administrative data includes the rate spread on each sampled mortgage, defined as the interest rate minus the Freddie Mac Primary Mortgage Market Survey average rate on a loan of the same type (fixed or adjustable) and term. The rate spread is top-coded at 1.5 percentage points, affecting 12 percent of consumers.[5] Yet the top-coding at 1.5 in NSMO is helpfully matched by bottom-coding at 1.5 in a separate federal dataset collected under the Home Mortgage Disclosure Act (HMDA)[6] Rate spreads in the HMDA data are based on the federally defined Average Prime Offer Rate (APOR). However, PMMS is the primary source used in constructing the APOR. The HMDA data uses the Annual Per-

[3] A full description all variables is available from FHFA at https://www.fhfa.gov/DataTools/Downloads/Documents/NSMO-Public-Use-Files/NSMO-Codebook-and-Tabulations-20190212.pdf

[4] https://www.marketwatch.com/story/a-farewell-to-arms-americans-still-shun-adjustable-rate-mortgages-10-years-after-the-crisis-2018-07-09

[5] The data is also bottom-coded at -1.5 percentage points, affecting 0.2 percent of the sample.

[6] The HMDA data is available at https://www.ffiec.gov/hmda/hmdaflat.htm.

centage Rate (APR) which differs from the interest rate by including discount points and fees.

I replace each top-coded spread in NSMO with a randomly selected spread between 1.5 and 3.7 from the 2016 HMDA data, where the upper bound of 3.7 was selected from the HMDA data as the 90th percentile of spreads of 1.5 or more (or the 99.2 percentile of all spreads). Observations are matched by whether the loan was for the purpose of refinancing, by which government agency (if any) insured the loan, and within bins of loan amount and income. This procedure helps to restore some of the variation in the data removed by top-coding. The drawbacks of this procedure are (1) the excess of rate spread over 1.5 is not linked to the full set of consumer characteristics and (2) the distribution of APR is used to replace part of the distribution of interest rate despite differences between the two rates.

There is a great deal of variation in the rate spread, with a standard deviation of 0.65 percentage points (or 0.56 percentage points if top-coded rate spreads are left as 1.5 percentage points). The rate spread depends to some extent on borrower and loan characteristics directly related to the cost and risk of lending. These include the borrower's credit score, the loan amount, the loan-to-value and debt-to-income ratios, and whether the loan is insured by by a government agency such as the Federal Housing Administration (FHA). I regress the rate spread on these credit characteristics.[7] The residual from this regression represents a cost-adjusted rate spread that attempts to exclude the interest rate effects of borrower-level variation in creditworthiness. Table 1.2 shows both the raw and adjusted rate spreads, with the calculation of the adjusted rate spread detailed in Appendix Table 1.A1. The R^2 from adjusting the rate spread is 0.07, indicating that most price differences in this market are not explained by the main measures of creditworthiness used by lenders.

While most consumers find a mortgage directly through a bank or other lender, 41 percent of consumers used a mortgage broker. Mortgage brokers are independent agents who assist the consumer in searching multiple lenders. During the 2013-2016 data period there was a significant change in the regulatory environment for brokers. The Consumer Financial Protection Bureau's (CFPB) Loan Originator Compensation rule was implemented, prohibiting payments by lenders to brokers in exchange for steering consumers to the lender's products. The rule was formally issued on January 20, 2013 and implemented January 10, 2014.

As shown in Table 1.2, consumers who ultimately accepted an offer through a broker searched only somewhat less than consumers who accepted an offer directly. This suggests that even with a broker, deliberate search may be necessary to find the best mortgage.

[7] In this linear regression, I account for top-coding using a Tobit model rather than by the replacement procedure outlined above.

Used broker	Non-recallers	Recallers	Difference
2013	1.09	0.30	0.79
2014-2016	1.36	0.28	1.08
Change	0.27	-0.01	0.29
No broker	Non-recallers	Recallers	Difference
2013	1.07	0.21	0.87
2014-2016	1.22	0.20	1.02
Change	0.15	0.00	0.15

Table 1.1: Mean cost-adjusted interest rate spread, by loan origination year and by whether the consumer used a mortgage broker in his shopping process. The CFPB Loan Originator Compensation Rule was implemented on January 10, 2014.

Consumers who used a broker tended to pay more than consumers who did not, which can be explained by less qualified consumers being more likely to need a broker's assistance in finding a loan. Among respondents who used a broker, the mean credit score was 724 compared to 733 for respondents who did not use a broker, a difference of 0.14 standard deviations.

The results of this paper are similar whether the analysis sample includes all brokered and unbrokered transactions (the main analysis), unbrokered transactions only, brokered transactions in 2013 (before the CFPB rule went into effect), or brokered transactions in 2014-2016. In particular, Table 1.1 shows no evidence that consumers using brokers were less likely to receive high rate spreads after the rule went into effect, or that the link between interest rate non-recall and the rate spread was weakened after the rule implementation.

An important limitation of NSMO and of many but not all other mortgage datasets is the omission of "discount points", which are upfront fees paid by the borrower in exchange for a lower interest rate. One discount point costs 1 percent of the loan amount and reduces the interest rate by some amount determined by the lender. Consumers may also pay negative discount points, increasing the interest rate in return for cash to pay closing costs. Bhutta and Hizmo (2019) find that an apparent tendency for racial minorities to pay higher interest rates is in fact fully explained by white borrowers paying more discount points. If some low interest rates in NSMO are in fact due to consumers paying points, this paper will tend to overestimate the benefit and cost of search and to underestimate price sensitivity. Similarly, the difference in interest rate outcomes between the consumer types discussed in the next section might be either attenuated or exacerbated by accounting for discount points. Bhutta et al. (2018) show however that points cannot explain much of the variation in mortgage outcomes. They find that adding discount points to a regression for rate spread reduces the

	Independent search	Used broker	Total
Share of sample	0.590	0.410	1.000
Rate spread	0.379	0.456	0.411
	(0.716)	(0.735)	(0.725)
Rate spread, adj. for characteristics	0.345	0.435	0.382
	(0.677)	(0.698)	(0.687)
# Lenders/brokers considered	1.735	1.640	1.696
	(0.850)	(0.794)	(0.829)
# Lenders/brokers applied to	1.276	1.289	1.281
	(0.572)	(0.606)	(0.586)
N	8,794	6,171	14,965

Table 1.2: Summary of the NSMO data. Standard deviations in parentheses. Observations are weighted to account for sampling design and survey non-response.

10th to 90th percentile range only from 0.60 to 0.54 percentage points.[8] The lowest rates in the data are likely to be especially contaminated by points paid. Section 1.3.2 discusses the treatment of outliers in the estimation.

The public version of the NSMO data omits data on lender identity. Alexandrov and Koulayev (2017) use NSMO with lender identity data to show that lenders differ substantially in their non-price quality. While I allow for nonprice variation in the quality of offers, the absence of lender identity prevents decomposing nonprice quality into a lender component and an idiosyncratic component as Alexandrov and Koulayev (2017) do.

Another lender-level feature of the mortgage market not modeled here is the incumbency advantage. Allen et al. (2014) and Allen et al. (2019) incorporate lender identity into a model of the (Canadian) mortgage market. In both papers the researchers observe the consumer's "home bank", i.e. the bank at which the consumer already has an account. Consumers begin by searching the home bank at a cost of zero and search other banks only if the home bank's offer is unsatisfactory. In Allen et al. (2014) the home bank is assumed to know the consumer's preferences, allowing the bank to make a just-acceptable initial offer to each of its existing customers. In Allen et al. (2019) the bank does not know consumer preferences. Dispersion in the distribution of search costs thus causes some consumers to reject the home bank's initial offer in favor of search. In this case the home bank may still win the consumer's business with a new offer, but this offer must now compete

[8]The 10th to 90th percentile range for my analysis sample is much wider at 1.6 pp since it does not incorporate lender fixed effects.

with offers from other searched firms. The model explains the tendency of banks with large retail market shares to offer less favorable rates to all customers, as well as the tendency of banks to price discriminate against their existing customers.

This paper presents a model with similar implications. Lenders price discriminate against consumers who appear less likely to search. However, this classification of consumers is made based on consumer inattention rather than loyalty. Given their higher search costs and lower price sensitivity, inattentive consumers have in a sense an indiscriminate loyalty to whatever high-priced offer is currently on the table. As in Allen et al. (2019), firms use this loyalty to extract a higher price.

1.2.1 Interest rate recall

The cost-adjusted rate spread has a standard deviation of 0.62 percentage points. For context, a fairly typical mortgage during this period was a 30-year fixed rate loan of $300,000 at a 4 percent APR. A loan one standard deviation above this APR at 4.65 percent would have a monthly payment of $1,547 instead of $1,432 and would cost fully $41,000 more over the 30-year term of the loan.

A first step toward identifying the causes of price dispersion is to search for some relation between the (unadjusted) rate spread and the approximately 400 survey and administrative variables reported in NSMO. With many potential explanatory variables and no theoretical basis for a specific functional form, I draw from the machine learning literature and fit a random forest model. The random forest model allows a convenient calculation of each variable's importance, defined as the increase in mean squared error caused by dropping the variable from the model. Appendix Table 1.A2 reviews the top 25 variables by importance. One variable is over 9 times as important as any other and accounts for over 40 percent of total variable importance. This variable is an indicator for whether the consumer responded to the question "What is the interest rate on this mortgage" by filling a a supplied blank space with a number, as opposed to leaving the field blank or checking a separate box for "Don't know". Neither the actual numerical response nor its correctness is available in the data. Recall of the interest rate is thus imperfectly measured.

Given the random forest results, I classify consumers into the two discrete types of interest rate recallers and non-recallers. This discrete type definition allows for convenient estimation of the structural model. Moreover, interest rate recall seems unlikely to be strongly related to the cost of lending after controlling for industry standard measures of creditworthiness. The same cannot be said for some other consumer characteristics, such as the consumer's self-reported initial level of concern about qualifying for a mortgage.

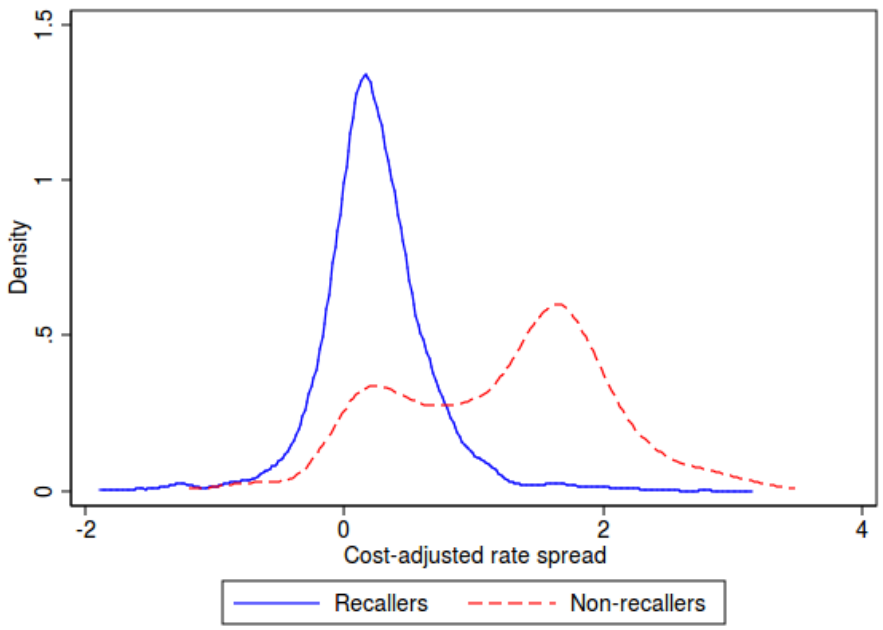

Figure 1.1: Density of cost-adjusted interest rate spreads paid, by whether the consumer recalled an interest rate on the survey (not necessarily the correct rate).

The 13.8 percent of consumers who were interest rate non-recallers paid substantially higher rate spreads, as shown in Table 1.3 and Figure 1.1. Compared to recallers, non-recallers were also less likely to have a firm idea of the mortgage they wanted and were more likely to choose a lender before choosing a loan type. Non-recallers did engage in slightly less search as measured by the number of firms seriously considered, but this difference is small relative to the difference in rate spread. Moreover, the next subsection will show that differences in the number of firms considered are not associated with large differences in rate spread.

One might argue that interest rate non-recall is a response to a high rate spread rather than a cause of it. Consumers who accepted a very high rate may have discovered their mistake by the time they complete the survey, in which case social desirability bias could motivate leaving the question blank or marking "Don't know". However, non-recallers also tend to report lower familiarity with all aspects of the market at the time they "began the process of getting this mortgage". Non-recallers were also less likely to recall other key loan characteristics such as the presence of a prepayment penalty or interest-only payments. These correlations suggest that interest rate non-recall is, at least for some consumers, part of an overall pattern of behavior. In addition, Table 1.A2 shows that other recall items such as recall of the monthly payment and amount borrowed were also associated with the interest rate spread. As with recall of the interest rate, consumers who recalled these items

	Non-recaller mean	Recaller mean	p-value	Overall mean
Rate spread	1.208	0.166	0.000	0.309
(std. dev.)	(0.807)	(0.481)		(0.725)
Cost-adjusted rate spread	1.250	0.244	0.000	0.382
(std. dev.)	(0.792)	(0.440)		(0.688)
Initially very familiar with rates	0.503	0.617	0.000	0.601
...with loan types available	0.396	0.475	0.000	0.464
Initially very familiar with process	0.487	0.522	0.005	0.518
Initial firm idea of loan wanted	0.538	0.605	0.000	0.596
Picked lender before loan type	0.733	0.685	0.000	0.692
# lenders/brokers applied to	1.241	1.317	0.000	1.307
(std. dev.)	(0.557)	(0.592)		(0.586)
...seriously considered	1.610	1.730	0.000	1.714
(std. dev.)	(0.799)	(0.834)		(0.829)
Loan purpose was refinance	0.407	0.414	0.571	0.413
Loan was through a broker	0.406	0.415	0.497	0.414
Recall: Had prepayment penalty	0.796	0.864	0.000	0.854
Recall: Balloon payment	0.819	0.881	0.000	0.872
Recall: Interest-only payments	0.798	0.873	0.000	0.863
Low APR very important (7,126)	0.784	0.808	0.083	0.805
Recalled closing costs (7,126)	0.126	0.458	0.000	0.416
Loan amount (approximate)	206.397	233.358	0.000	229.639
(std. dev.)	(119.6)	(130.9)		(129.8)
Respondent non-Hispanic white	0.800	0.775	0.014	0.779
...non-Hispanic black	0.058	0.050	0.141	0.051
...non-Hispanic Asian	0.031	0.053	0.000	0.050
...non-Hispanic other race	0.027	0.027	0.945	0.027
...Hispanic	0.083	0.095	0.123	0.093
...female	0.514	0.428	0.000	0.440
Multiple borrowers	0.468	0.464	0.740	0.465
Observations	2,394	12,571		14,965

Table 1.3: Summary statistics by whether the consumer recalled an interest rate (not necessarily the correct rate). The p-values are for a t-test of whether non-recallers and recallers have the same mean. Some questions were only asked in certain survey waves, in which case the sample size is noted. Loan amounts are reported only categorically and are therefore approximate.

paid lower rates.

Another predictor of paying a high interest rate may be failure to return the survey. The NSMO public file includes only consumers who returned the survey and assigns each consumer an analysis weight to account for the sampling rate of each wave and for non-response. One of the "key predictive variables" in the non-response adjustment is the rate spread, and indeed a rate spread of 1.5 percentage points or more is associated with a 36 percent higher analysis weight. The NSMO public file does not include actual records for non-responders, and naturally the survey variables such as number of firms considered will be unknown for this group. For these reasons I do not include an indicator for survey response in the analysis. However, the association between survey non-response and high rate spreads provides additional evidence that the consumer's underlying level of conscientiousness affects his or her success in the market.

Table 1.3 also tabulates the respondent's race and gender by interest rate recall. Several papers have investigated the possibility of race or gender discrimination in the mortgage market. Bartlett et al. (2018) finds that black and Hispanic borrowers pay a small interest rate penalty of 0.06-0.09 percentage points after controlling for creditworthiness and loan characteristics. However, Bhutta and Hizmo (2019) find that these racial disparities are explained by white borrowers choosing to pay more discount points. Using data on subprime mortgages, Fang and Munneke (2017) find that female sole borrowers pay an interest rate 0.13 percentage points higher than joint male and female borrowers, even after controlling for the higher loan termination risk of female sole borrowers.

Table 1.3 shows that survey respondents who do not recall the interest rate are not significantly more likely to be black or Hispanic but are significantly more likely to be female. Table 1.4 further shows that women are less likely to recall the interest rate regardless of whether they were responding for a household of multiple borrowers or for themselves only. These findings suggest that the gender disparity found by Fang and Munneke (2017) may be caused by statistical discrimination against less attentive consumers. In this paper's data, loans with a female respondent have an increased cost-adjusted rate spread of 0.05 percentage points, as shown in Tables 1.5 (for sole borrowers) and 1.6 (for women responding on behalf of multiple borrowers). Column (2) of these tables shows that the gender disparity becomes statistically insignificant once interest rate recall is controlled for.

Interestingly, Table 1.6 shows that the disparity against female respondents persists when there are multiple borrowers–in most cases, a male-female couple. One explanation is that women who respond to the survey on behalf of a couple or family likely took the lead in mortgage negotiations, incurring roughly the same penalty as female sole borrowers.

Higher interest rates thus apply to women as negotiators rather than as borrowers.

Since attention as measured by interest rate recall is correlated with gender, the simple discriminatory strategy of charging higher rates to inattentive consumers would have a disparate impact and may be prohibited by law. However, firms could modify this strategy to include some compensating discrimination in favor of women. Absent a race or gender element, charging higher prices to inattentive consumers is legally permitted.

Moreover, discrimination by consumer attentiveness does not necessarily require a lender to "size up" each consumer for an individually tailored offer. Discrimination could be implemented simply by leaving an opening for negotiation that inattentive consumers might not exploit.[9] The structural model of Section 1.3 assumes that lenders may price discriminate by consumer attentiveness as represented by interest rate recall, but cannot price discriminate by any other consumer characteristic.

	Male	Female
One borrower	0.123	0.157
Multiple borrowers	0.115	0.165

Table 1.4: Share of survey respondents who do not recall the interest rate, by respondent gender and whether there were multiple borrowers on the mortgage.

1.2.2 Data on search intensity

The structural model of Section 1.3 requires data on the number of firms a consumer searched. NSMO asks the consumer "How many different lenders/mortgage brokers did you seriously consider before choosing where to apply for this mortgage?", and subsequently how many lenders/brokers the consumer actually applied to. According to their self-reporting, 48 percent of consumers seriously considered only one firm, 35 percent considered two, and 16 percent considered three or more. This finding is consistent with past survey evidence. Cai and Shahdad (2015) find that 35 percent of consumers obtain only one mortgage quote. Woodward and Hall (2012) cite surveys showing that the modal number of loans considered was two (Lacko and Pappalardo, 1991) and that more than half of borrowers considered only one loan (Federal Reserve Board, 2009), which reinforces their structural estimate of a common search intensity of two.

[9] Some states impose a fiduciary duty on mortgage brokers, though not on mortgage lenders. Fiduciary duties require the broker to seek the best loan for the borrower and thus lessen the opportunity for any sort of discrimination. In addition, the CFPB Loan Originator Compensation rule that took effect during the data period further limited the scope of broker price discrimination by preventing lenders from compensating brokers who steer business to the lender.

	(1)	(2)
	Cost-adjusted rate spread	Cost-adjusted rate spread
Respondent is non-Hispanic black	0.0457	0.0636
	(0.0512)	(0.0438)
Respondent is non-Hispanic Asian	-0.0284	-0.0440
	(0.104)	(0.0893)
Respondent is Hispanic	0.0970*	0.125***
	(0.0439)	(0.0376)
Other race	-0.133	-0.0425
	(0.0898)	(0.0770)
Respondent is female	0.0515	0.0177
	(0.0331)	(0.0284)
Recalled interest rate		-0.952***
		(0.0409)
Constant	0.343***	1.170***
	(0.0248)	(0.0414)
Observations	1502	1502
Adjusted R^2	0.004	0.269

Standard errors in parentheses

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Table 1.5: Regression of the cost-adjusted rate spread on demographic characteristics of the respondent. Sample: Loans with only one borrower.

	(1)	(2)
	Cost-adjusted rate spread	Cost-adjusted rate spread
Respondent is non-Hispanic black	-0.0578	-0.00900
	(0.0985)	(0.0842)
Respondent is non-Hispanic Asian	-0.240	-0.142
	(0.129)	(0.110)
Respondent is Hispanic	0.00792	0.0606
	(0.0591)	(0.0506)
Other race	0.0557	0.0252
	(0.153)	(0.131)
Respondent is female	0.0256	-0.0122
	(0.0411)	(0.0352)
Recalled interest rate		-0.952***
		(0.0501)
Constant	0.343***	1.170***
	(0.0332)	(0.0519)
Observations	981	981
Adjusted R^2	-0.001	0.270

Standard errors in parentheses

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Table 1.6: Regression of the cost-adjusted rate spread on demographic characteristics of the respondent. Sample: Loans with more than one borrower.

This paper labels as search intensity the number of firms seriously considered, whose mean of 1.71 is consistent with the surveys cited above. Two possible types of error may arise. First, there may be substantial noise in the reported number of firms seriously considered, including noise caused by different interpretations of the survey question. I discuss the implications of these idiosyncratic errors for a simultaneous search model in subsection 1.3.1. However, idiosyncratic errors will not affect the sequential search model at the center of this paper, under which the rate spread is not correlated with the search intensity of the individual consumer (as opposed to the consumer's type). A greater concern is systematic error. Search intensity may not be an appropriate proxy for the full array of consumer search behavior. Consumers may have a great deal of information on other firms without going so far as to seriously consider them. As Deltas and Li (2018) note for their measure of search in the mortgage market (the loan application to loan origination ratio), search intensity is "a proxy for less formal rate queries and other information acquisition efforts." As do Woodward and Hall (2012), this paper finds that using consumer-reported search intensity as the model's formal search intensity leads to reasonable structural estimates. I also follow Woodward and Hall (2012) in presenting results under alternative values of search intensity as discussed in section 1.6.

Table 1.7 regresses the cost-adjusted interest rate on indicators for the number of firms seriously considered. Consumers who consider multiple firms do obtain lower rate spreads, but the effect is less than 0.03 percentage points.

1.3 Model

Each consumer has a type: whether or not the consumer was among the "attentive" 86 percent of consumers who recalled an interest rate when prompted. I assume that this type is observable to firms. If consumer i of type m accepts the offer of firm j and searches n_{im} times, his utility is

$$u_{im} = v_{ijm} - c_m n_{im} \tag{1.1}$$

where

$$v_{ijm} = -\alpha_m p_{jm} + \varepsilon_{ijm} \tag{1.2}$$

is consumer i's utility of the accepted offer, consisting of a price disutility $-\alpha_m p_{jm}$ and

	(1)	(2)
	Adj. spread (non-recallers)	Adj. spread (recallers)
Num. lenders/brokers seriously considered: 1	0	0
	(.)	(.)
2	-0.0965*	-0.0215*
	(0.0433)	(0.00944)
3	-0.0645	-0.0281*
	(0.0647)	(0.0140)
4	0.383***	-0.0196
	(0.106)	(0.0291)
5	-0.0100	-0.00230
	(0.137)	(0.0504)
Constant	1.285***	0.255***
	(0.0269)	(0.00615)
Observations	2394	12571
Adjusted R^2	0.004	0.000

Standard errors in parentheses

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Table 1.7: Regression of cost-adjusted rate spread on the number of lenders seriously considered. Column (1) includes only consumers who did not recall their interest rate, column (2) includes only those who did.

an idiosyncratic nonprice utility $\varepsilon_{ij} \sim \text{Gumbel}(0,1)$ iid. Note that there is no consumer subscript on the search cost c_m. Past work that has identified a full distribution of search costs (Hortacsu and Syverson, 2004) has typically had access to additional data such as firm identities and the distribution of offered prices (as opposed to accepted prices).

The price p_{jm} is operationalized as the cost-adjusted rate spread. If firm j has cost r_{jm} of serving a consumer of type m, it will set its price p_{jm} to type m consumers to maximize

$$\max_{p_{jm}} \Pr(i \text{ accepts } p_{jm})(p_{jm} - r_{jm}) \tag{1.3}$$

.

I assume that before they search, consumers know the distribution of v_{ijm} for offers to their type but have no information on the v_{ijm} they would receive from any particular firm. Consumers thus engage in undirected search and all consumers draw firms from the same sampling process. I defer until subsection 1.3.1 determining whether search is sequential or simultaneous.

I assume that ε_{ijm} is independent of the firm's cost r_{jm} and is unobserved by the firm. This makes ε_{ijm} independent of the offered price p_{jm}. The assumptions on ε are not testable in this data, though some potential biases can be qualified. To the extent that firms with high costs and high rate offers provide better service with higher nonprice quality, observed consumer behavior will be as if α were smaller than its true value. To the extent that firms observe their match value ε_{ijm} and tailor their price offers by charging more when ε_{ijm} is high, observed consumer behavior will again be as if α were smaller than its true value. However, it is difficult to qualify the overall effect of these potential biases on the sequential search model estimation of Subsection 1.3.2.

In the final stage of the sequential search estimation, I assume that both consumer types face the same distribution of firm costs ($F_0^{rO} = F_1^{rO}$). Estimation involves choosing the model parameters to equate these distributions as closely as possible. Recall that prices and firm costs are in terms of the interest rate spread adjusted for conventionally measured creditworthiness in Table 1.A1. The assumption $F_0^{rO} = F_1^{rO}$ thus rules out only cost differences not explained by differences in credit score, income, or other standard credit covariates.

Nonetheless, the assumption does rule out an facially plausible alternative explanation of the high rate spreads paid by non-recallers. Instead of price discriminating, firms may simply face a higher credit risk in lending to non-recallers and may pass this cost through in the form of higher rate spreads. Section 1.6 presents evidence against this cost passthrough

story. In particular, the interest rate disparity between recallers and non-recallers persists in the market for FHA-insured mortgages where the government bears all credit risk.

1.3.1 Simultaneous search implies implausibly weak price preferences

Consumers may choose in advance how many firms to search (simultaneous search) or may decide after each search whether to continue searching (sequential search). Which search model is more appropriate will depend on the market in question. For example, Moraga-González et al. (2015) argue that search in the Dutch automobile market is simultaneous given the need to arrange test drives in advance. Honka and Chintagunta (2017) formally estimate which model of search applies to the US auto insurance industry and find evidence for simultaneous search. In the mortgage market Woodward and Hall (2012) adopt simultaneous search while Deltas and Li (2018) are agnostic about whether search is sequential, simultaneous, or a process that combines elements of both. This section discusses the data patterns that suggest sequential search as the more plausible data generating process for the NSMO data.

Simultaneous search implies a rigid relationship between search intensity and the price paid, which can be simulated. Consider for now only the more numerous of the two consumer types, those who recall interest rates. Of these consumers, 47 percent searched once and 36 percent searched twice. Recall that conditional on consumer type, all consumers are assumed to draw from the same distribution of offered prices. Under simultaneous search this offer distribution can be directly observed as the distribution of accepted prices for $n_i = 1$. Given this offer distribution, a simulation exercise finds the unique value of price sensitivity α that explains the observed mean price paid by $n_i = 2$ consumers of the same consumer type.

The simulation procedure is presented in detail in Appendix 1.A. The resulting estimate is $\alpha_m = 0.21$ for recallers, implying that one standard deviation of offered price has the same value as 0.070 standard deviations of the Gumbel$(0, 1)$ offered non-price utility.[10] This is an implausibly low utility value for nearly half a percentage point of rate spread. Further, 80 percent of consumers who applied to more than one lender reported "searching for better loan terms" as one of their reasons for doing so, which is inconsistent with search being mainly for nonprice characteristics.

A separate reason to avoid a simultaneous search model stems from the supply side and the substantial share of consumers who search only once. If almost half of consumers

[10]Repeating this exercise for non-recallers gives $\alpha_m = 0.12$, implying that one standard deviation of the non-recaller offered price distribution has the same utility value as 0.073 standard deviations of offered non-price utility.

really are precommitted to searching only one firm then a firm's optimal offer may be an extremely high interest rate that sets consumer surplus to zero. Since this is of course not the market reality, firms must believe that their consumers could search elsewhere even if they do not. This belief is the basis of the sequential search model.[11]

1.3.2 Estimating the sequential search model

Given the evidence in the previous subsection, I assume that search is sequential. This section presents the estimation method in three stages. The first two stages consider only one consumer type (recallers or non-recallers). First, accepted prices and the average number of firms searched are used to estimate w_m conditional on an assumed value of α_m. Second, α_m and w_m together are used to estimate the firm cost corresponding to each accepted price. I reweight the empirical cost distribution corresponding to accepted prices to find its counterpart for offered prices. Together these two stages give one estimated firm cost distribution F_0^{rO} for each α_0 in a given parameter range to explore, and similarly F_0^{rO} for each α_1 in another range.

Third, I apply the assumption that firms can serve each consumer type at the same cost. The cost distribution for each type's offered prices should thus be the same. I search over (α_0, α_1) and compare the cost distributions found by the first two stages. The pair (α_0, α_1) with the most similar cost distributions is accepted as the estimate, from which the first two estimation stages recover (w_0, w_1). The remainder of this subsection discusses each stage in detail.

1.3.2.1 Stage 1: Estimate w_m conditional on α_m

The first stage estimates w_m conditional on α_m for a single type. All consumers within the type have the same search cost and thus the same strategy, which is the reservation utility rule of Weitzman (1979). A consumer of type m accepts any offer that yields utility of at least w_m, where w_m is defined implicitly by

$$c_m \equiv \int_{w_m}^{\infty} (x - w_m) f^{vO}(x; m) dx \tag{1.4}$$

where $f^{vO}(\cdot; m)$ is the density of offered utility to type m and $f^{pO}(\cdot; m)$ will be used

[11] In their simultaneous search model Woodward and Hall (2012) must omit the option of searching only once, for the same reason: "if the borrower continues to believe that there is, in effect, only one broker in the universe, the broker can capture up to the total benefit to the borrower from buying *any* house." (emphasis in original)

to represent the same for offered price.[12] Since neither the offer distribution nor the consumer's decision rule depends on how many searches the consumer has already made, the distribution of accepted price (or utility) is independent of n_i. Recall from Table 1.7 that while accepted prices in the data are decreasing in the number of searches, the relation is very weak with only a 0.02 percentage point benefit for searching twice instead of once.

The probability that an offer is accepted is

$$\Pr(\text{accept}|p,m) = \Pr(-\alpha_m p + \varepsilon_{ijm} > w_m) \tag{1.7}$$
$$= \Pr(\varepsilon_{ijm} > w_m + \alpha_m p) \tag{1.8}$$
$$= 1 - F_\varepsilon(w_m + \alpha_m p) \tag{1.9}$$

The density of offered prices to consumer type m is not directly observed, but it can be recovered conditional on α_m from the consumer type's density of accepted prices and and their mean search intensity. Let p_{im} be the accepted price for consumer i of type m. The distribution of accepted price is given by

$$\Pr(p_{im} < p) = \Pr(p_{jm} < p|\text{accept}) \tag{1.10}$$
$$= \frac{1}{\Pr(\text{accept})} \Pr(p_{jm} < p, \text{accept}) \tag{1.11}$$
$$= \mathrm{E}[n_{im}|m] \Pr(p_{jm} < p, \text{accept}) \tag{1.12}$$
$$= \mathrm{E}[n_{im}|m] \int_{-\infty}^{\infty} \Pr(p_{jm} < p, \text{accept}|p_{jm} = p^*) f_m^{pO}(p^*) dp^* \tag{1.13}$$
$$= \mathrm{E}[n_{im}|m] \int_{-\infty}^{p} \Pr(\text{accept}|p^*, m) f^{pO}(p^*|m) dp^* \tag{1.14}$$

.

Differentiate with respect to p, replacing $\mathrm{E}[n_{im}|m]$ with its sample analogue $\overline{n}_m$ and $\Pr(\text{accept}|p^*, m)$ with its value from Equation 1.9.

[12] Formally,

$$f^{vO}(x;m) = \frac{d}{dx} Pr(-\alpha_m p_{jm} + \varepsilon_{ijm} < x) \tag{1.5}$$
$$f^{pO}(x;m) = \frac{d}{dx} Pr(p_{jm} < x) \tag{1.6}$$

$$f^{pA}(p;m) = \overline{n}_m(1 - F_\varepsilon(w_m + \alpha_m p))f^{pO}(p;m) \tag{1.15}$$

$$f^{pO}(p;m) = \frac{f^{pA}(p;m)}{\overline{n}_m(1 - F_\varepsilon(w_m + \alpha_m p))} \tag{1.16}$$

.

The density $f^{pO}(p;m)$ must integrate to 1:

$$1 = \int_{-\infty}^{\infty} \frac{f^{pA}(p;m)}{\overline{n}_m 1 - F_\varepsilon(w_m + \alpha_m p)} dp \tag{1.17}$$

.

This integral is approximated by treating accepted prices as simulation draws from the density $f^{pA}(\cdot;t)$.

$$1 \approx \sum_{i:m_i=m} wt^A_{ijm} \frac{1}{\overline{n}_m(1 - F_\varepsilon(w + \alpha p))} \tag{1.18}$$

where wt^A are the survey weights reported in NSMO.

Fixing α_m and w_m allows evaluating the acceptance probability of Equation 1.9 and thus evaluating the RHS of Equation 1.18. The RHS of 1.18 is decreasing in each price's acceptance probability, and each price's acceptance probability is decreasing in α_m and w_m. Equation 1.18 thus implicitly yields one of type m's unknown parameters (suppose α_m) in terms of the other (w_m).

1.3.2.2 Stage 2: Recover the firm cost distribution F^{rO}_m conditional on α_m

The second estimation stage recovers the distribution of firm costs conditional on the demand parameters α_m and w_m. Suppose firm j has cost r_{jm} and encounters a consumer i known to be of type m. The firm's expected profit with price p is

$$\pi_{ijm} = (p - r_{jm})(1 - F_\varepsilon(w_m + \alpha_m p)) \tag{1.19}$$

.

The first order condition is

$$0 = (1 - F_\varepsilon(w_m + \alpha_m p)) - \alpha_m(p - r_{jm})f_\varepsilon(w_m + \alpha_m p) \tag{1.20}$$

leading to the optimal markup

$$(p - r_{jm}) = \frac{1}{\alpha_m} \frac{(1 - F_\varepsilon(w_m + \alpha_m p))}{f_\varepsilon(w_m + \alpha_m p)} \tag{1.21}$$

.

Equation 1.21 is used to recover the cost corresponding to each offer. To find the distribution of firm costs, I compute weights that account for the original NSMO sampling and non-response weights as well as the average number of offers for each acceptance at a given price. The overall weight is

$$wt^O_{ijm} = wt^A_{ijm} \frac{1}{1 - F_\varepsilon(w_m + \alpha_m p_{jm})} \tag{1.22}$$

.

Weighting the estimated r_{jm} by wt^O_{ijm} gives a representative sample from the firm cost distribution F^{rO}_m to consumer type m. With only one consumer type this would be as far as the estimation strategy could go, providing only a set-identification result (which Subsection 1.4.1 presents). The third estimation stage achieves full identification under the additional assumption that the same firms serve two observably different consumer types, who differ in their preferences but not in their distribution of cost of lending after adjusting for conventionally measured creditworthiness.

1.3.2.3 Stage 3: Equate (approximately) F^{rO}_0 and F^{rO}_1

The third and final estimation stage seeks to make the firm cost distributions as similar as possible across the two consumer types. I assume that once the interest rate is adjusted for observable cost and risk factors in Section 1.2, interest rate recall is unrelated to the cost of lending. For each α_m, I compute the implied w_m for each consumer type using Equation 1.18. I use the markup rule 1.21 to recover firm costs and the reweighting 1.22 to estimate the distribution of firm costs corresponding to offers. Finally, I use a grid search to select an α_m for each type that minimizes the distance between the two estimated cost distributions.

The cost distributions are compared using the Earth Mover's Distance (EMD) of optimal transport theory (Rubner et al., 1998). Intuitively, suppose that each cost distribution is a collection of piles of dirt at the locations given by the estimated costs and with the size of each pile given by the corresponding weight from Equation 1.22. The EMD is the amount of work, in terms of mass times distance, required to transform one distribution into

the other. Since the lowest rate spreads in NSMO would imply extremely low firm costs, I limit sensitivity to outliers by omitting the bottom 25 percent of each cost distribution. Appendix Table 1.A5 and Appendix Figure 1.A2 repeat the estimation dropping only the bottom 10 percent of each cost distribution. The 25 percent drop was selected as the base specification because its parameter estimates implied fewer implausibly low costs. The lowest observed interest rates are likely explained by consumers paying discount points unobserved in the data, as discussed in section 1.2.

1.4 Results

1.4.1 Interest rate recallers only

Figure 1.2 shows the estimated price offer density in terms of the undetermined price coefficient α, as well as the corresponding density of costs corresponding to these offers. Only the 86 percent of consumers who recalled the interest rate are included.[13] Consideration of multiple consumer types is deferred to the next section.

At $\alpha = 0$ consumers are indifferent to price and the price offer density is identical to the density of accepted prices. Repeated search is only for the purpose of finding a better non-price utility. At the other extreme a high α of 2 or 3 implies that there are many offers at very high prices. Intuitively, a high price sensitivity implies a low probability of accepting a high price, which means that the few accepted high prices are a small fraction of a multitude of high price offers.

[13] Appendix Figure 1.A1 presents the price offer and cost densities for non-recallers.

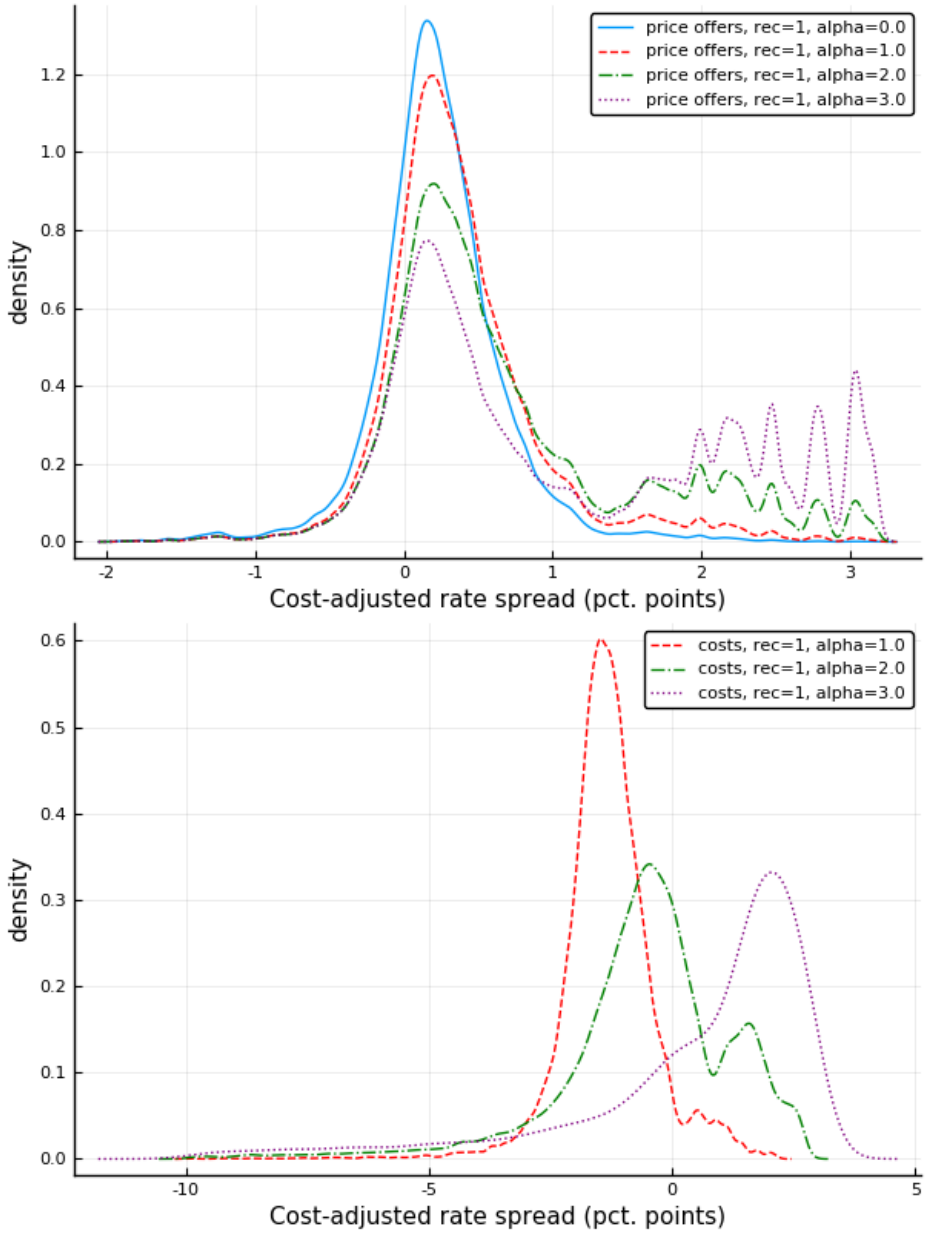

Figure 1.2: Possible probability densities of price offers and firm costs, based on data from consumers who recalled the interest rate. Selecting the correct density requires identifying the price disutility α using data from both consumer types. There is no cost density for $\alpha = 0$ because the optimal price to a completely price-insensitive consumer is infinite.

1.4.2 Full results using both consumer types

Consumers who did not recall their interest rate paid much higher cost-adjusted rate spreads (1.34 versus 0.17 percentage points) and searched only slightly less, seriously considering an average of 1.60 firms instead of 1.72. Sequential search does not impose a direct link

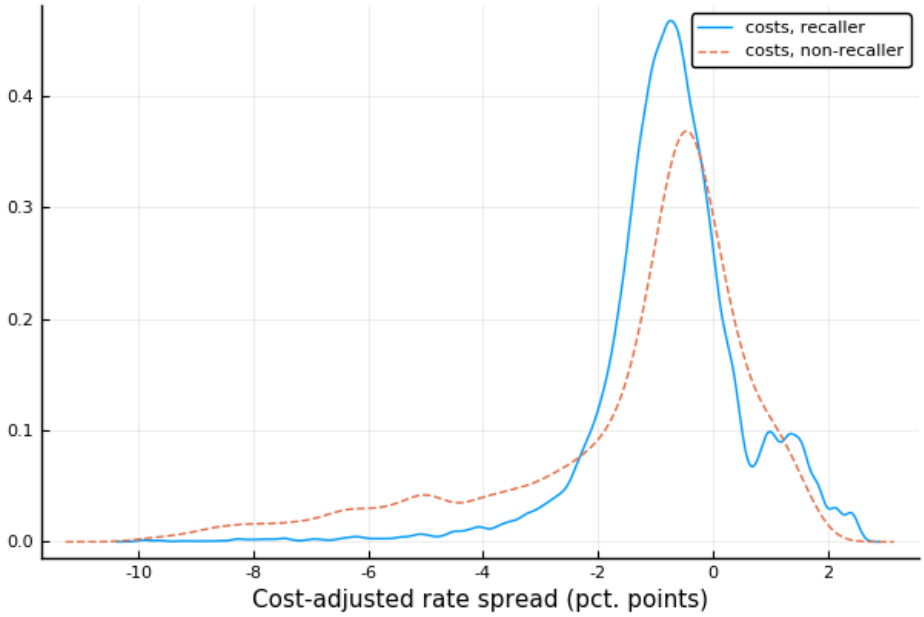

Figure 1.3: Density of firm costs, corresponding to the offer distribution to each consumer type.

between the number of searches and the price paid as would be seen with simultaneous search. However, it is possible that the two types of consumers had different preferences α_m, w_m and that this caused lenders and brokers to provide them with different price offer distributions. This section presents estimates under the assumption that consumer types may differ in preferences but not in the costs of the firms they searched.

Table 1.8 presents the estimation results and Figure 1.3 shows the implied cost densities. Recallers are much more price sensitive by any metric. They have higher α_m, a higher ratio of the utility value of one standard deviation of searched price to the utility value of one standard deviation of searched nonprice utility ε, and a lower price threshold at which the probability of accepting an offer equals $\frac{1}{2}$. Figure 1.3 shows that the cost densities are only a rough match at the estimated parameters, which shows that there is some misspecification in this model. In particular the non-recaller cost density is has more mass at the extremes than the recaller cost density. This suggests that the tendency of *some* non-recallers to receive extremely high offers cannot be fully explained by *all* non-recallers having uniformly different tastes or higher search costs. A model with heterogeneity in search costs within a type might reduce the misspecification but may also be more difficult to estimate reliably.

While the estimation strategy in general does not guarantee a unique distance-minimizing (α_0, α_1), identification is clear with the current data. Figure 1.5 shows that the EMD objective descends monotonically toward a unique minimum and that there are no near-minima at very different parameters.

	Nonrecallers	Recallers
α	0.71	1.57
	(0.03)	(0.12)
c	0.84	0.8
	(0.07)	(0.08)
w	-1.03	-0.57
	(0.02)	(0.02)
$\frac{c}{\alpha}$	1.19	0.51
	(0.04)	(0.03)
$\frac{\Delta u(\text{1 std. dev. accepted price})}{\Delta u(\text{1 std. dev.}\varepsilon)}$	0.45	0.53
	(0.03)	(0.09)
Price such that $\Pr(\text{accept}) = 0.5$	1.96	0.59
	(0.13)	(0.04)
Q1 of firm cost distribution F_{rg}^O	-2.19	-1.33
	(0.1)	(0.09)
Q2 of firm cost distribution F_{rg}^O	-0.7	-0.74
	(0.12)	(0.12)
Q3 of firm cost distribution F_{rg}^O	-0.07	-0.11
	(0.11)	(0.09)
IQR of firm cost distribution F_{rg}^O	2.12	1.22
	(0.03)	(0.01)

Table 1.8: Demand estimates: α_m for recallers and non-recallers, the implied value of the reservation price parameter w, and selected other statistics. The parameters α_m were chosen to minimize the Earth Mover's distance between recaller and non-recaller cost densities. Standard errors (in parentheses) generated by 1,000 bootstrap replications.

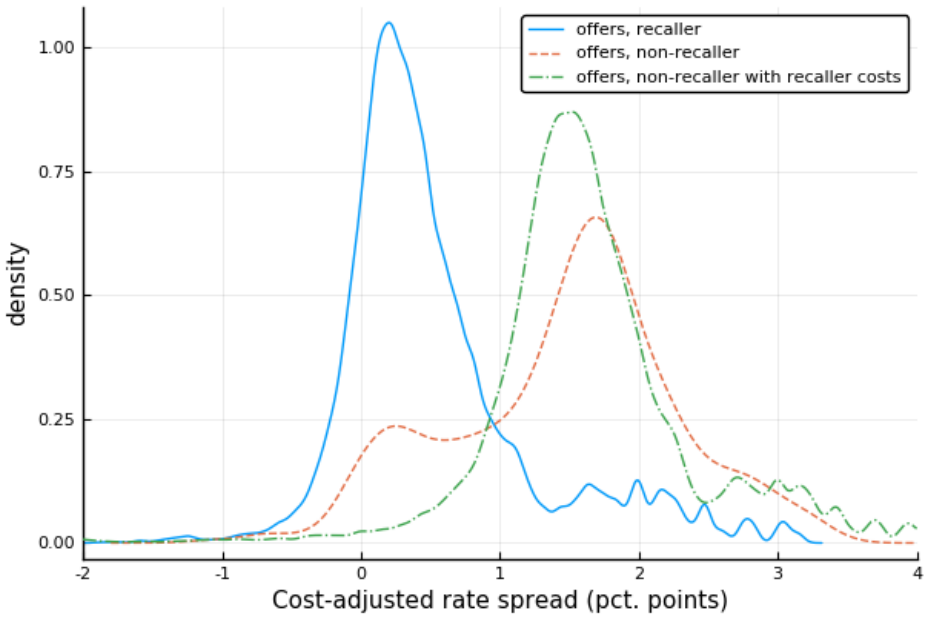

Figure 1.4: Density of price offers, by consumer type. The solid line shows the offer distribution that non-recallers would have faced if their distribution of lending costs had been exactly that of recallers rather than approximately the same as in Figure 1.3.

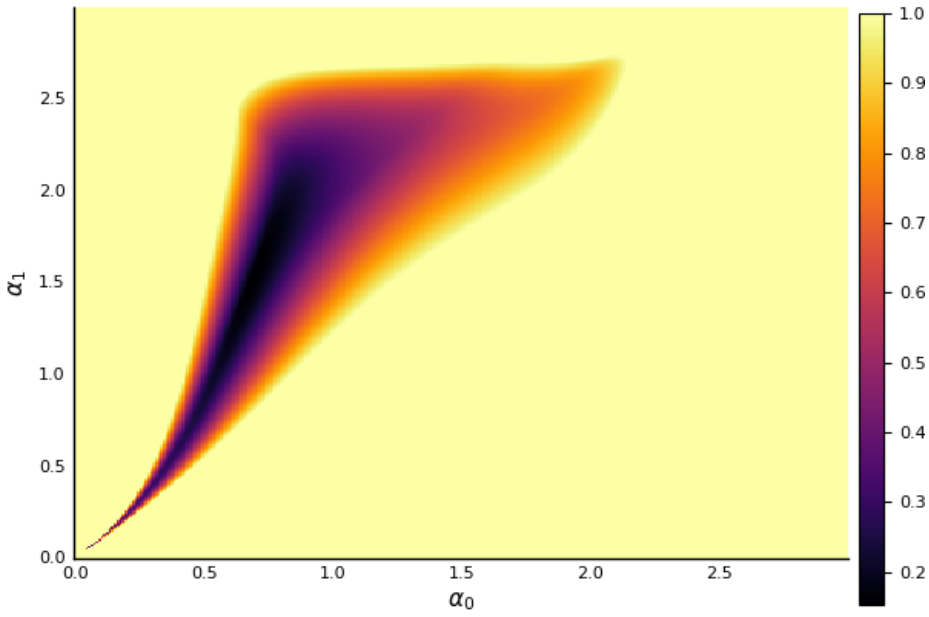

Figure 1.5: Earth Mover's Distance as a function of the price sensitivities α_0, α_1 of non-recallers and recallers respectively. For visualization, the distance is top-coded at 1.

1.5 Counterfactuals

Consumers who do not recall their interest rate pay a large penalty in risk-adjusted rate spread: 1.01 percentage points. This paper's structural model allows a decomposition of this penalty by its several causes, as shown in Table 1.9. First, non-recallers would pay 0.24 percentage points less solely due to having a slightly lower distribution of lending costs than recallers. Since the estimation strategy is based on matching the two lending cost distributions, this discrepancy represents the degree to which the model fails to fit the data. Possible causes include unobserved differences between the two consumer types such as patterns of discount points paid. With the lending cost distribution for non-recallers held fixed at that of recallers, the disparity to be explained grows to 1.25 percentage points. This is the difference between the (actual) mean price of 0.24 for recallers and the (hypothetical) mean price of 1.49 for non-recallers with the same cost of lending distribution as recallers.

To decompose this disparity, first suppose that an individual non-recaller had the preferences and search behavior of a recaller while continuing to face price discrimination. Such a consumer would pay a mean price of 1.34, as shown in the last row of Table 1.9. Thus 0.15 percentage points (1.49 minus 1.34) of the non-recaller difference is attributable directly to differences in behavior. Next make the opposite assumption that an individual non-recaller managed to pose as a recaller and avoid price discrimination. The consumer continues to search and choose according to non-recaller preferences. Such a consumer pays a mean price of 0.37, as shown in the second-to-last row of the table. Therefore the disparity attributable to price discrimination is 1.49 minus 0.37, or 1.12 percentage points. The remaining 0.02 percentage points of the disparity is an interaction between differences in the preferences of the two types and differences in their offer distributions.

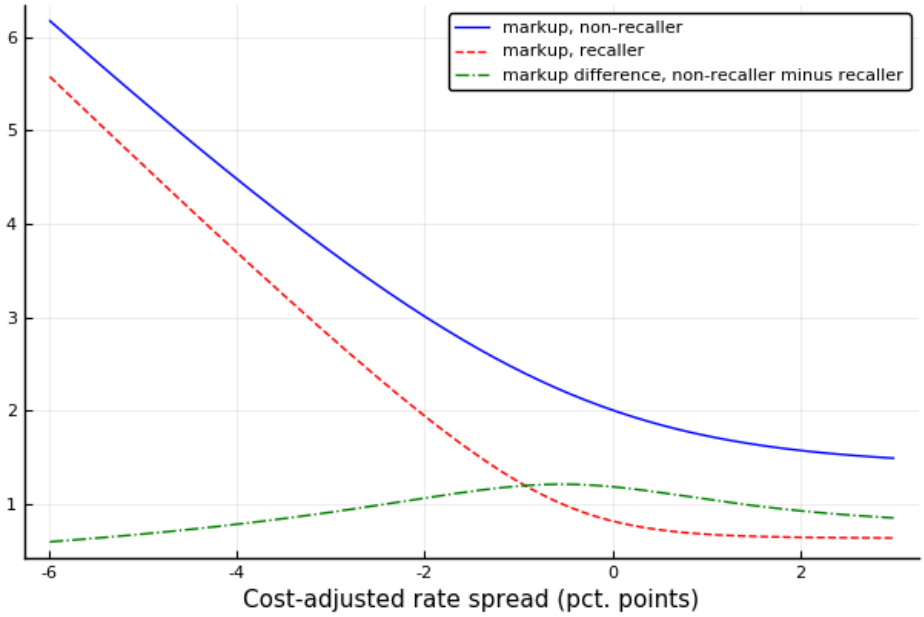

Figure 1.6: Firm's optimal markup as a function of the cost of lending. Based on Equation 1.21.

	Mean accepted price	Mean num. searches
Recallers	0.24	1.73
Non-recallers	1.23	1.61
Non-recaller preferences with predicted non-recaller offers	1.46	1.59
Non-recaller preferences with recaller offers	0.36	1.56
Recaller preferences with predicted non-recaller offers	1.31	1.84

Table 1.9: Results of counterfactually altering preferences and offers. "Predicted non-recaller offers" in the third row are those that would be made by price-discriminating firms to non-recallers if the cost of lending were exactly as for recallers rather than just approximately the same.

The large price discrimination effect reflects the very different optimal markups to the two consumer types. Figure 1.6 shows the result of solving for the optimal markup to each type using Equation 1.21. For both consumer types, markups are largest when costs are low. The markup is always higher for non-recallers, with the largest disparity occurring at moderate to high costs. Referring to Figure 1.4, this difference in markups causes non-recallers to often face high offers that would be very rare for recallers.

The counterfactual results are largely unaffected by whether the data includes trans-

actions with or without brokers, before or after the CFPB Loan Originator Compensation rule implementation. An open question is why using a broker offers no protection against the negative effects of interest rate non-recall. Given the CFPB rule, brokers differ from lenders in that they have no incentive to deliberately offer worse rates to less attentive or conscientious consumers. One possibility is that brokers take consumer preferences into account when searching on their behalf, and the preferences of non-recallers are relatively price-insensitive. Alternatively some level of attention may be required of the consumer even when assisted by a broker.

1.6 Features of the mortgage market and robustness

This section pauses to consider several possible features of the mortgage market that could affect this paper's results. First, consumer search attempts may be undercounted. The definition of search most appropriate to an economic model may not be the same as the definition of search as number of lenders "seriously consider[ed]". A consumer may have some level of knowledge about a firm and its offers from viewing advertising or from some form of casual search not rising to the point of serious consideration.

The estimation results in Appendix Tables 1.A3 and 1.A4 are analogous to Tables 1.8 and 1.9 but with search intensity doubled for all consumers. While the results are numerically quite different, the same qualitative conclusions apply. Results appear to depend on the small size of the recaller/non-recaller difference in search intensity rather than on the overall level of search in the market.

A second and opposite possibility is that search attempts are overcounted, particularly for the least creditworthy borrowers. Agarwal et al. (2017) show that consumers who search more actually pay higher rates, counter to the prediction of a standard search model. In their model this pattern is explained by the presence of differences in consumer creditworthiness. Less creditworthy consumers will receive higher rate offers and must search more to find a lender willing to accept them. To address this concern, Appendix Table 1.A6 repeats the analysis for the most creditworthy borrowers only and finds similar results.

Perhaps the most important fact of the mortgage market not captured in this paper's data is the presence of discount points. Bhutta and Hizmo (2019) analyze data on points and interest rate spreads for FHA-insured loans. The mean number of points paid was negative, meaning that consumers were accepting higher interest rates in exchange for cash to pay closing costs. Consumers who accepted higher rate spreads received on average some compensation in the form of extra points. A consumer in the tenth decile of rate spread received on average 1.63 more points than a consumer in the first decile while paying about

an interest rate about 1.1 percentage points higher.

Receiving extra points dampens the negative impact of paying a high interest rate. However, this compensation is far from complete. Continuing with the example of a 30-year fixed rate mortgage of $300,000, suppose that the market average interest rate was 4 percent. The results of Bhutta and Hizmo (2019) suggest that a consumer in the tenth decile of rate spread would pay an interest rate around 4.8 percent compared to 3.7 percent for a consumer in the first decile. The high interest rate borrower would have a monthly payment of $1,574 versus $1,381 for the low interest rate borrower, an excess of $193 per month. After 26 months the higher payment would have wiped out the initial points received, with almost 28 years of higher payments remaining on the mortgage. For a borrower discounting at 4 percent annually, the difference in present value of the two monthly payment streams is approximately $40,000 in favor of the low interest rate. Even for a borrower discounting rapidly at 12 percent annually, the difference is approximately $19,000. Despite receiving $4,890 worth of points, the high interest rate borrower is far worse off–unless perhaps he refinances quickly into a lower interest rate mortgage. Although there are no prepayment penalties in the FHA market studied by Bhutta and Hizmo (2019), closing costs for a refinance are typically around 1.5 percent of the loan amount and would thus eliminate most or all of the benefit of receiving points on the first loan.

These calculations suggest that much price dispersion remains after accounting for points, and that this paper's exclusive focus on interest rates does not grossly overstate price dispersion. The structural estimation tends to discard those consumers who paid the most points by excluding from the distance calculation the bottom 25 percent of the lending cost distribution for each consumer type.

On the other hand, Bhutta and Hizmo (2019) use secondary market data to show that high interest rate loans are not more profitable for the lender. This presents a puzzle: the rate-points tradeoff is immaterial for the lender but appears very consequential for the consumer. In the extreme case that points do fully compensate the consumer for interest rate changes, this paper in effect shows a tendency of inattentive consumers to pay higher rates versus higher points. The result would not rule out dispersion in the overall price of loans, but would be unrelated to such dispersion. Data that combines consumer characteristics, interest rates, and points paid would be useful for investigating this issue.

1.6.1 Loan performance and FHA insurance

Another reason why non-recallers pay more may be a higher default or prepayment risk. A lender charging high rates to non-recallers may simply be passing through the higher

cost of lending rather than engaging in price discrimination. The loan performance data in NSMO allows for a limited test of this explanation. Indicators for delinquency or default are observed for one month in each quarter from the date of the loan through the end of my data period in June 2018. Each loan thus has at least 18 months of performance data.

Central to any assessment of a lender's risk is whether the loan is insured by the government. In the case of a loan insured by the Federal Housing Administration (FHA) the government rather than the lender bears the risk of default in order to encourage the lender to offer mortgages to certain higher-risk borrowers. [14] Tables 1.10 and 1.11 show that both for "conventional" loans without a government guarantee and for FHA-insured loans, non-recall of the interest rate is associated with roughly a doubling of the delinquency and serious delinquency rates. Table 1.12 shows that the interest rate penalty for non-recall is similar for FHA and conventional loans.

While the higher delinquency and default rates for non-recallers are consistent with lenders passing through higher costs, the result of Table 1.12 is not. If the interest rate penalty for non-recall were due to the passthrough of credit risk then one would expect to see no interest rate penalty for non-recall in the FHA market, where risk is borne by the government. Instead the penalty is of similar size in both markets. On the other hand, price discrimination by search intensity is an explanation equally applicable to both the conventional and FHA markets and is consistent with the similar patterns observed in the two markets.

[14] In the event of foreclosure on a conventional loan, the lender takes possession of the property. If instead the loan is FHA-insured, the government takes possession of the property and pays the lender the entire remaining amount owed on the loan. The FHA's assumption of all default risk for these loans is key to the analysis strategy of Woodward and Hall (2012).

Conventional:

	(1) ever_delin	(2) ever_90_delin	(3) ever_default
Recalled interest rate	-0.371	-0.293	-2.800**
	(0.283)	(0.612)	(1.051)
Adjusted rate spread	-0.0189	-0.315	-0.619*
	(0.156)	(0.315)	(0.298)
Predicted rate spread			4.464***
			(1.048)
Creditworthiness controls	Yes	Yes	No
Observations	10365	8272	10365
Pseudo R^2	0.226	0.302	0.145

Standard errors in parentheses

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

FHA:

	(1) ever_delin	(2) ever_90_delin	(3) ever_default
Recalled interest rate	-0.853***	-1.029*	-1.924*
	(0.259)	(0.416)	(0.980)
Adjusted rate spread	-0.224	-0.342	-1.232***
	(0.159)	(0.239)	(0.372)
Predicted rate spread			2.771
			(1.526)
Creditworthiness controls	Yes	Yes	No
Observations	2464	2464	2483
Pseudo R^2	0.128	0.187	0.066

Standard errors in parentheses

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Table 1.10: Regression of delinquency or default on rate recall and the cost-adjusted rate spread. The dependent variables are indicators for ever having any delinquency on the loan, any delinquency more than 90 days, and any foreclosure or default.

	Conventional, Non-recallers	Conventional, Recallers	FHA, Non-recallers	FHA, Recallers
ever_delin	0.0303	0.0157	0.141	0.0774
ever_90_delin	0.00655	0.00386	0.0539	0.0271
ever_fail	0.00296	0.000242	0.0132	0.00475
N	1,691	8,674	392	2,091

Table 1.11: Delinquency rates by FHA versus conventional loans and recall of the interest rate. Variables are indicators for any ever having delinquency on the loan, any delinquency more than 90 days, and any foreclosure or default.

Conventional:

	(1) Cost-adjusted rate spread
Recalled rate	-0.999***
	(0.0235)
Constant	1.237***
	(0.0230)
Observations	10365
Adjusted R^2	0.329

Standard errors in parentheses

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

FHA:

	(1) Cost-adjusted rate spread
Recalled rate	-0.950***
	(0.0449)
Constant	1.190***
	(0.0435)
Observations	2483
Adjusted R^2	0.267

Standard errors in parentheses

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Table 1.12: Robustness check on FHA loans. Relation of interest rate recall to the cost-adjusted interest rate spread, by FHA loans versus loans without a government guarantee.

1.6.2 Relation to models of rational inattention

Price dispersion may alternatively be explained by a rational inattention model rather than a search model. As proposed by Matejka and McKay (2015), a consumer may be able to choose any firm in the market while being uncertain of the value of choosing each firm. Instead of searching a discrete set of firms, the consumer engages in a more abstract information processing strategy to gather signals about the alternatives. Matejka and McKay (2015) show that under certain assumptions[15], the probability of choosing alternative j is

$$\Pr(i \text{ chooses } j|m) = \frac{e^{v_{ijm}/\lambda_m}}{\sum_q e^{v_{iqm}/\lambda_m}} \tag{1.23}$$

where λ_m is the unit cost of information processing for consumer type m. It is apparent from Equation 1.23 that data on accepted prices cannot identify both λ_m and the preference for unobserved nonprice product characteristics. An increase in either quantity would amount to reduced price sensitivity and would have the effect of translating a given number of observed high accepted prices into a larger number of high offered prices. Therefore this section treats price as the only distinguishing characteristic of a mortgage offer, so that λ_m in practice captures both information frictions and nonprice preferences.

Appendix 1.B shows that given the choice probabilities in Equation 1.23, a small firm with cost r_j would choose price $p_{jm} = r_j + \lambda_m$. Assuming that the two consumer types share the same distribution of r_j, the distribution of accepted prices should be the same for each type except for a location shift due to different λ_m. The simplest estimation method would be to estimate $\lambda_0 - \lambda_1$ as $\mathrm{E}[p_{ijm}|m=0] - \mathrm{E}[p_{ijm}|m=1]$, which equals 1.01.[16] Non-recallers would thus be estimated to have a higher cost of processing information. Still, it would not be possible to identify the absolute level of λ_0 or λ_1.

In the data however, one accepted price distribution is not simply a shifted version of the other. Figure 1.1 shows that the accepted price distributions are differently shaped for each consumer type. Unlike with the sequential search model, it is impossible to find parameters that give both types approximately the same distribution of cost of lending.

The combination of the rational inattention model and the common costs assumption

[15] Specifically, the information processing cost follows a specific entropy-based form and the alternatives are *a priori* homogeneous.

[16] To be precise, we would like to estimate the difference between mean offered prices rather than mean accepted prices. Appendix 1.B presents a preliminary discussion of how to calculate offer weights that relate the unobserved price offer distributions to the observed accepted price distributions. This calculation is analogous to Equation 1.22 in the main structural model of Section 1.3. Overall, offered prices will tend to exceed accepted prices more for the attentive consumer type than for the inattentive. This will tend to cause the difference $\lambda_0 - \lambda_1$ to be smaller than the mean difference in accepted prices.

is thus rejected by the data. As both this paper and Alexandrov and Koulayev (2017) find, brand preferences or other nonprice preferences play a large role in consumers' mortgage decisions. Subsuming nonprice preferences into the same parameter as the information friction produces a simple model that is grounded in the behavioral economics literature but that overlooks a key feature of the market, which may contribute to its apparent misspecification. Data that includes lender identifiers may improve the performance of this model by disentangling two possible reasons for not choosing the lowest interest rate: unawareness that the rate was available versus a desire to prioritize nonprice aspects of the mortgage.

1.7 Conclusion

Interest rate spreads in the mortgage market are far more variable than can be explained by variation in the cost and risk of lending according to conventional metrics. This paper identifies failure to recall the interest rate as a strong predictor of paying a higher rate. The 14 percent of consumers who did not recall their interest rates paid a full percentage point of additional interest after controlling for cost-related variables. Correlations between interest rate non-recall and other measures of inattention suggest that interest rate non-recall is proxying for broad naïveté or disengagement from the process.

Consumer search is limited to about 1.7 firms on average, consistent with past survey evidence. The weak relation between search intensity and price paid suggests that a sequential search model fits the market better than a simultaneous model of the sort used by Woodward and Hall (2012). Estimating the sequential search model, I find that differences in interest rate outcomes between recallers and non-recallers cannot be explained by differences in search intensity and are thus caused by differences in the optimal stopping rule and in the distribution of offers received. Point identification rests on the assumption that interest rate recall is unrelated to the cost of lending after controlling for conventional metrics such as credit score and the loan-to-value ratio.

This paper assumes that both consumer types search rationally using a sequential search model, allowing for both price and nonprice preferences. Assuming that lenders have the same distribution of costs when lending to both types, the data is best explained by the two types having moderately different preferences leading to substantial price discrimination.

Counterfactuals establish that about 10 percent of the penalty for interest rate non-recall arises from differences in search strategy, with the remaining 90 percent coming from differences in the estimated distribution of price offers. This result has important implications for the optimal consumer search strategy and for consumer education. Price

insensitivity and high search costs in themselves do not cause much worse outcomes for inattentive consumers. However, having such preferences causes lenders to optimally price discriminate heavily against the inattentive consumer. This discrimination could take the form of "sizing up" the consumer and making an individually tailored offer, but it may be more simply and realistically accomplished by leaving an opening for negotiation that inattentive consumers fail to exploit.

In light of the substantial price discrimination found in this paper, consumer education should continue to advocate negotiating to obtain the best possible offer as opposed to simply searching more firms. The CFPB's "Owning a Home" guide for instance suggests contacting at least three lenders and negotiating on price, using loan estimates from competing lenders to improve the consumer's bargaining power. The strong link between interest rate non-recall and paying a high rate suggests that motivating and informing the least sophisticated mortgage borrowers could substantially reduce their tendency to pay high rates.

Finally, it may be objected that the stronger price preference of recallers suggests they may be paying more discount points or settling for worse non-price attributes. This would imply that the consumer welfare penalty for non-recall is smaller than the price penalty would suggest. Non-recallers do appear less sensitive to price when choosing when to stop search, but it is not clear whether this is a true preference or simply the result of ignorance that there are lower prices available. Data that includes discount points paid and lender identifiers would be useful to determine whether recallers traded other loan attributes for a lower rate as opposed to finding a better deal in both dimensions.

Appendix

1.A Estimating Price Sensitivity α_m under Simultaneous Search

This section presents the simulation procedure used to estimate α_m for a single consumer type. Consumers of a given type (suppose recallers) are divided into those who search once ($n_i = 1$) and those who search twice ($n_i = 2$), with all higher search intensities discarded.[17] Under simultaneous search the distribution of accepted prices for $n_i = 1$ is equal to the offer distribution.

I consider a large number of identical hypothetical $n_i = 2$ consumers and simulate two offers to each. Each offer consists of a rate spread p_{jm} sampled from the prices accepted by $n_i = 1$ consumers and a nonprice utility ε_{ijm} sampled from the assumed Gumbel$(0, 1)$ distribution. For each consumer there is a critical value α_i^* above which the consumer switches from the higher to the lower priced of his two offers. It is possible that α_i^* is negative. For some consumers the two prices are identical, in which case α_i^* is undefined.

For consumers with two distinct prices I compute α_i^* by

$$-\alpha_i^* p_{i1m} + \varepsilon_{i1m} = -\alpha_i^* p_{i2m} + \varepsilon_{i2m} \tag{1.24}$$

$$\alpha_i^*(p_{i1m} - p_{i2m}) = \varepsilon_{i1m} - \varepsilon_{i2m} \tag{1.25}$$

$$\alpha_i^* = \frac{\varepsilon_{i1m} - \varepsilon_{i2m}}{p_{i1m} - p_{i2m}} \tag{1.26}$$

For a given α, all consumers with $\alpha_i^* < \alpha$ choose the higher priced of their two offers and all consumers with $\alpha_i^* > \alpha$ choose the lower priced. This implies that the simulated mean accepted price for $n_i = 2$ consumers is a weakly decreasing function of α. There

[17] As shown in Table 1.7, searching more than two firms is not associated with a substantial further decrease in price paid. Using $n_i > 2$ consumers in this simulation exercise would imply an even weaker price sensitivity.

	(1) Rate spread
Loan-to-value LTV (percent)	-0.00146 (0.0113)
LTV^2	0.0000121 (0.0000656)
Combined loan-to-value CLTV (percent)	0.00362 (0.0113)
CLTV^2	0.00000104 (0.0000650)
Payment-to-Income PTI (percent)	0.00313 (0.00170)
PTI^2	-0.0000391 (0.0000253)
Debt-to-Income DTI (Percent)	0.00296 (0.00155)
DTI^2	-0.0000125 (0.0000172)
Multiple borrowers	-0.190 (1.181)
First mortgage in credit file	-0.0448** (0.0138)
Jumbo loan	-0.172*** (0.0289)
VantageScore 3.0 at Origination	-0.00723*** (0.00182)
Score^2	0.00000451*** (0.00000127)
Coborrower score minus 750	-0.000663*** (0.000172)
(Coborrower score minus 750)^2	-0.00000126 (0.00000224)
Freddie Mac PMMS rate	-1.192 (0.643)
PMMS^2	0.0875 (0.0828)
Refinance	0.0674*** (0.0121)
Year-quarter FEs	Yes
Loan type FEs (e.g. FHA, conventional)	Yes
Loan amount category FEs	Yes
Income category FEs	Yes
Constant	6.489*** (1.937)
σ	0.564*** (0.00535)
Observations	14,965
Pseudo R^2	0.0680

Standard errors in parentheses
* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Table 1.A1: Tobit regression to adjust the rate spread for factors affecting the cost or risk of lending. σ is the standard error of the residual.

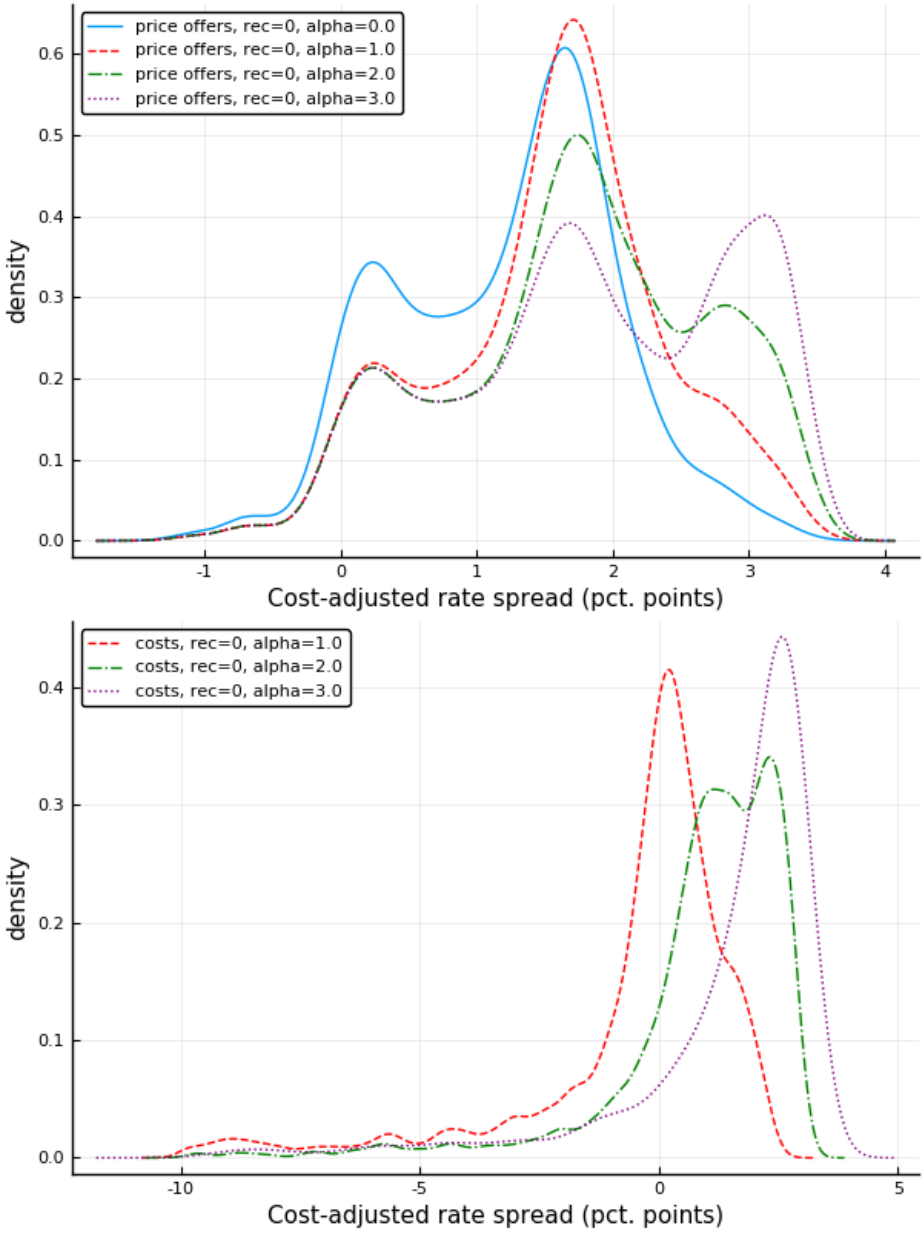

Figure 1.A1: Possible probability densities of price offers and firm costs, based on data from consumers who did not recall the interest rate. There is no cost density for $\alpha = 0$ because the optimal price to a completely price-insensitive consumer is infinite.

	NSMO variable	Variable description	Share of importance
1	z43	Recalled interest rate	0.415343
2	enterprise	Fannie/Freddie/no GSE	0.051097
3	yq	Year and quarter of loan	0.049234
4	z42	Recalled monthly payment	0.046592
5	z41	Recalled amount borrowed	0.024769
6	score_r	Respondent credit score	0.022335
7	loan_amount_cat	Loan amount (categorical)	0.018983
8	loan_type	Loan type, e.g. FHA	0.016593
9	pmms	PMMS average rate	0.01392
10	x66	Use of property	0.011471
11	x27b	Believes rate was lowest available	0.008424
12	cltv	Combined loan to value	0.007856
13	ltv	Loan to value	0.0078
14	dti	Debt to income	0.006722
15	z58	Recalled property price	0.005453
16	approx_survey_lag	Months from loan to survey wave	0.005434
17	z67	Recalled move-in date	0.005301
18	pti	Payment to income	0.005172
19	x74r	Respondent age (categorical)	0.004978
20	x76r	Respondent education	0.004573
21	survey_wave	Survey wave	0.004343
22	z64	Recalled rent received (if rental)	0.0041
23	x63	Rents out property	0.00371
24	x05a	Initial familiarity with rates	0.003707
25	score_s	Coborrower credit score	0.003606

Table 1.A2: The 25 most important variables in a random forest model of the cost-adjusted rate spread. Full survey questions and tabulations are available at https://www.fhfa.gov/DataTools/Downloads/Documents/NSMO-Public-Use-Files/NSMO-Codebook-and-Tabulations-20190212.pdf.

	Non-recallers	Recallers
α	0.53	1.21
c	0.35	0.32
w	0.25	0.62
$\frac{c}{\alpha}$	0.65	0.27
$\frac{\Delta u(\text{1 std. dev. accepted price})}{\Delta u(\text{1 std. dev.}\varepsilon)}$	0.33	0.41
Price such that $\Pr(\text{accept}) = 0.5$	0.22	-0.21
Q1 of firm cost distribution F^O_{rg}	-1.38	-0.92
Q2 of firm cost distribution F^O_{rg}	-0.63	-0.63
Q3 of firm cost distribution F^O_{rg}	-0.21	-0.23
IQR of firm cost distribution F^O_{rg}	1.17	0.69
Mean of top 75 pct of cost distribution	-0.34	-0.26

Table 1.A3: Robustness check: All search intensities doubled

	Non-recallers	Recallers
α	0.53	1.21
c	0.35	0.32
w	0.25	0.62
$\frac{c}{\alpha}$	0.65	0.27
$\frac{\Delta u(\text{1 std. dev. accepted price})}{\Delta u(\text{1 std. dev.}\varepsilon)}$	0.33	0.41
Price such that $\Pr(\text{accept}) = 0.5$	0.22	-0.21
Q1 of firm cost distribution F_{rg}^O	-1.38	-0.92
Q2 of firm cost distribution F_{rg}^O	-0.63	-0.63
Q3 of firm cost distribution F_{rg}^O	-0.21	-0.23
IQR of firm cost distribution F_{rg}^O	1.17	0.69
Mean of top 75 pct of cost distribution	-0.34	-0.26

Table 1.A4: Robustness check: Counterfactuals with all search intensities doubled.

	Non-recallers	Recallers
α	0.87	2.27
c	0.88	0.99
w	-1.31	-1.3
$\frac{c}{\alpha}$	1.01	0.44
$\frac{\Delta u(\text{1 std. dev. accepted price})}{\Delta u(\text{1 std. dev.}\varepsilon)}$	0.55	0.76
Price such that $\Pr(\text{accept}) = 0.5$	1.93	0.73
Q1 of firm cost distribution F_{rg}^O	-1.81	-1.95
Q2 of firm cost distribution F_{rg}^O	-0.26	-0.39
Q3 of firm cost distribution F_{rg}^O	0.4	0.92
IQR of firm cost distribution F_{rg}^O	2.21	2.87
Mean of top 75 pct of cost distribution	-0.54	-0.36

Table 1.A5: Robustness check: Alternative trimming of the cost distribution (1). Replicates Table 1.8 while minimizing the difference between cost distributions disregarding the bottom 10 percent of each distribution. The main estimation disregards the bottom 25 percent of each distribution.

	Mean accepted price	Mean num. searches
Recallers	0.24	1.71
Non-recallers	1.28	1.69
Non-recaller preferences with predicted non-recaller offers	1.37	1.72
Non-recaller preferences with recaller offers	0.4	1.63
Recaller preferences with predicted non-recaller offers	1.11	2.13

Table 1.A6: Robustness check: Limit to the most creditworthy borrowers. Replicates Table 1.9 using borrowers with no more than 80 percent loan to value and 18 percent payment to income ratios, credit score at least 772, and reporting that they were "not at all" concerned about qualifying for a loan. Among these consumers there are 1435 recallers and 256 non-recallers.

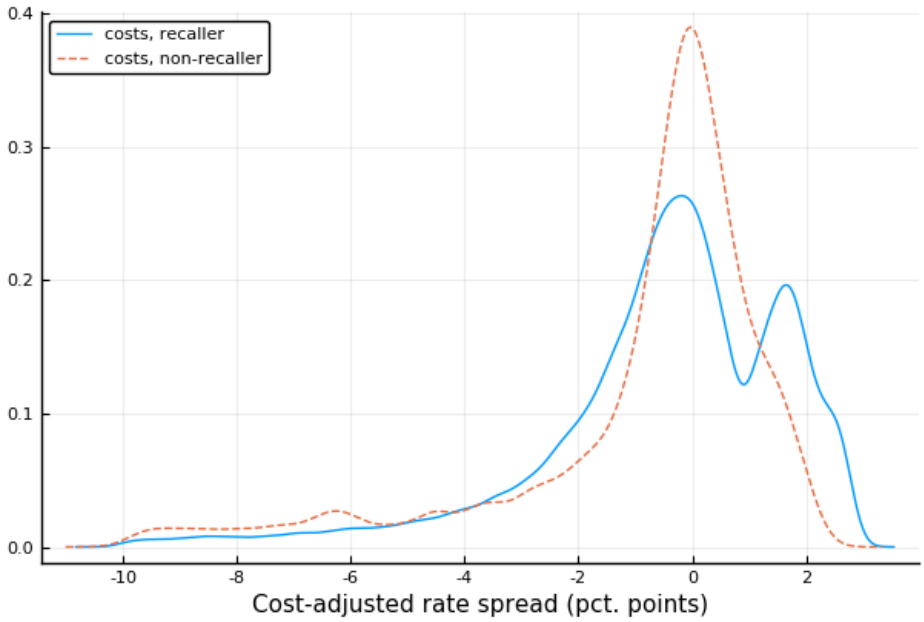

Figure 1.A2: Robustness check: Alternative trimming of the cost distribution (2). Replicates Figure 1.3 while minimizing the difference between cost distributions disregarding the bottom 10 percent of each distribution. The main estimation disregards the bottom 25 percent of each distribution.

is a range of α such that the simulated mean accepted price for $n_i = 2$ consumers is approximately equal to its sample analogue.

Mechanically, α is found as follows. Order the consumers by increasing α_i^*, treating undefined as $+\infty$. Compute (A) the cumulative sum of consumers' higher prices and (B) the reverse cumulative sum of consumers' lower prices. Then when α is just above α_i^*, the sum of accepted prices is the ith element of (A) plus the $N - i$ element of (B), where N is the number of simulated consumers. The implied mean accepted price is this value divided by N. I select the α that most closely matches the implied mean accepted price to to the sample analogue $E[p_{jm}|n_i = 2, m]$. With a large number of simulated consumers this match is essentially exact.

To demonstrate the method, Figure 1.A1 shows 8 simulated consumers instead of 10 million. Since consumers differ in the prices and nonprice utilities of their offers, each consumer has a different critical value α_i^* above which he chooses the lower priced of his two offers. Ordering the consumers by increasing α_i^*, consumer 1 has offers $L = (p_L = -0.14, \varepsilon_L = 1.35)$ and $H = (p_H = 0.12, \varepsilon_H = -0.76)$. Since the higher priced offer H is also worse in nonprice terms, it will only be chosen if the consumer has a strong taste for paying higher prices. Specifically, consumer 1's choice as a function of α is

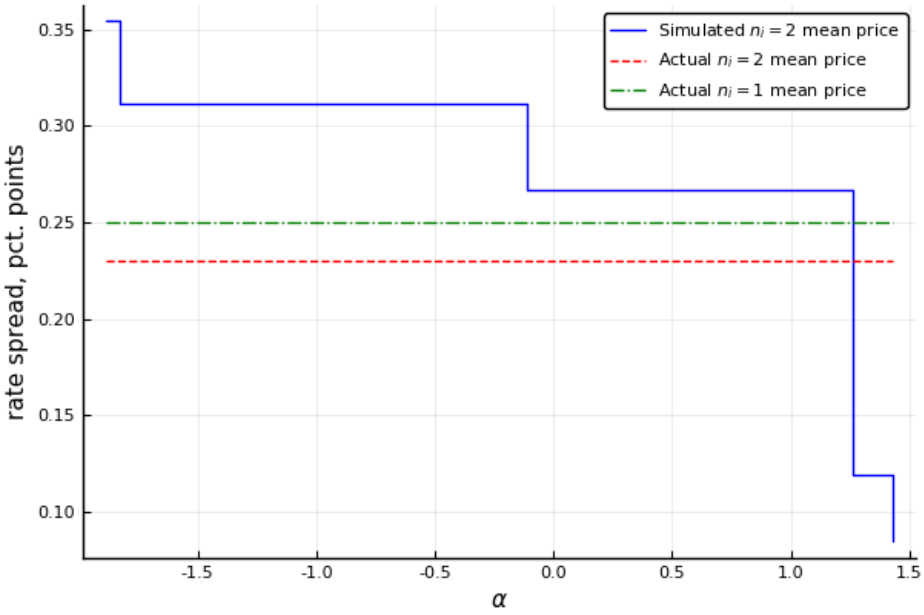

Figure 1.A1: Illustration of the simultaneous search model estimation method, using 8 simulated consumers instead of the 10 million used in the actual analysis. The estimate of α is the range $(-0.11, 1.26)$ (with $\alpha < 0$ implying that consumers like to pay more). Data: consumers who recalled the interest rate and seriously considered either 1 or 2 lenders/brokers.

$$\text{choose L} \iff \alpha > \alpha_1^*, \tag{1.27}$$

$$\alpha_1^* = \frac{\varepsilon_H - \varepsilon_L}{p_H - p_L} = -2.25 \tag{1.28}$$

For $\alpha < \alpha_1^*$ consumer 1 and all others accept the higher priced of their two offers, leading to a mean accepted price of 0.50. Once α increases above α_1^*, consumer 1 switches to his lower priced offer and the mean accepted price falls to 0.37. This process repeats at $\alpha_2^* = -1.88$, $\alpha_3^* = -1.83$ and so on as more consumers switch. For $\alpha \in (-0.11, 1.26)$ the mean accepted price is 0.27, which of the possible mean accepted prices best approximates the sample mean accepted price 0.23 for interest rate recalling consumers with $n_i = 2$. In this small-scale demonstration, $\alpha \in (-0.11, 1.26)$ would be the estimated price sensitivity needed to explain the data by a simultaneous search model.

Figure 1.A2 repeats this analysis with 10 million simulated consumers. The size of each step is now trivially small so that α is essentially point-estimated. Again, the estimated α is where the simulated mean accepted price equals the observed mean for $n_i = 2$, in this case $\alpha = 0.23$. Observe that where $\alpha = 0$ the simulated mean price instead equals the observed mean price for $n_i = 1$, which is the mean of the offer distribution. Intuitively,

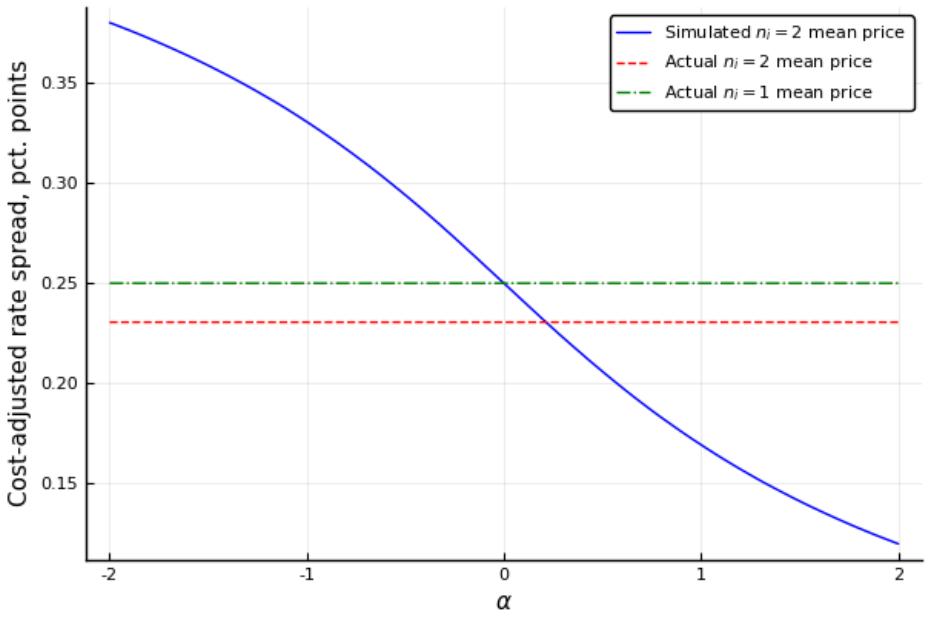

Figure 1.A2: Estimation of the sequential search model with 10 million simulated consumers. The estimated α is 0.21. Data: consumers who recalled the interest rate and seriously considered either 1 or 2 lenders/brokers.

$\alpha = 0$ makes consumers equally likely to choose the low price or the high, so that drawing $n_i = 2$ instead of $n_i = 1$ does not affect the distribution of accepted prices.

Matching predicted and observed $n_i = 2$ accepted prices with this procedure yields $\alpha = 0.21$. This value can be transformed into a relative utility of price versus non-price characteristics. The standard deviation of $n_i = 1$ accepted prices is 0.43, while the standard deviation of nonprice utility was normalized to 1.28 by the Gumbel$(0, 1)$ form. Each standard deviation of offered p is thus worth only $\frac{0.24 \cdot 0.43}{1.28} = 0.07$ of a standard deviation in offered ε.

1.B Derivation of the Markup under Rational Inattention

The choice probability 1.23 was developed by Matejka and McKay (2015) for a discrete choice set. I assume that the set of offers is discrete but that there are many offers in the market at each price. Specifically, suppose that there are N offers in the market and wt^O_{km} is the probability that an offer has price p_k. Then the probability of a consumer finding and accepting a specific offer with price p_k is

$$\Pr(i \text{ chooses } p_k \text{ from firm } j) = \frac{e^{-p_k/\lambda_m}}{\sum_q N w t_q^O e^{-p_q/\lambda_m}} \tag{1.29}$$

while the probability of accepting *any* offer with price p_k is

$$\Pr(i \text{ chooses } p_k) = \frac{N w t_{km}^O e^{-p_k/\lambda_m}}{\sum_q N w t_q^O e^{-p_q/\lambda_m}} \tag{1.30}$$

$$= \frac{w t_{km}^O \; A \; e^{-p_k/\lambda_m}}{A \sum_q w t_q^O e^{-p_q/\lambda_m}} \tag{1.31}$$

for any scalar A. Choose $A = \frac{1}{\sum_q w t_q^O e^{-p_q/\lambda_m}}$ to obtain

$$\Pr(i \text{ chooses } p_k) = w t_{km}^O \; A \; e^{-p_k/\lambda_m} \tag{1.32}$$

$$w t_{km}^O \; A = \frac{e^{-p_k/\lambda_m}}{\Pr(i \text{ chooses } p_k)} \tag{1.33}$$

.

The offer weights wt_q must sum to 1, implying

$$1 = \sum_q w t_q^O \tag{1.34}$$

$$A = \sum_q w t_q^O \; A \tag{1.35}$$

.

Use this value for A in Equation 1.33 to obtain

$$w t_{km}^O \sum_q w t_q^O \; A = \frac{e^{-p_k/\lambda_m}}{\Pr(i \text{ chooses } p_k)} \tag{1.36}$$

.

Applying 1.33 again gives

$$wt^O_{km} \sum_q \frac{e^{-p_q/\lambda_m}}{\Pr(i \text{ chooses } p_q)} = \frac{e^{-p_k/\lambda_m}}{\Pr(i \text{ chooses } p_k)} \tag{1.37}$$

$$wt^O_{km} = \frac{\frac{e^{-p_k/\lambda_m}}{\Pr(i \text{ chooses } p_k)}}{\sum_q \frac{e^{-p_q/\lambda_m}}{\Pr(i \text{ chooses } p_q)}} \tag{1.38}$$

recalling that $\Pr(i \text{ chooses } p_k)$ is estimated by the number of observed prices p_k weighted by their original NSMO survey weights wt^A_{ijm}. Equation 1.38 thus has only two unknowns for a given price p_k: the information processing cost λ_m and the weight wt^O_{km}. Thus given λ_m it is possible to identify wt^O_{km}, the frequency with which a price p_k is offered to type m.

Meanwhile on the supply side, the firm's problem can be written as

$$\max_p (p - r)e^{-p/\lambda_m} \tag{1.39}$$

using Equation 1.29 and the assumption that there are many firms (so the effect of one firm's price choice on the denominator of 1.29 is negligible). The firm's first order condition is

$$0 = e^{-p/\lambda_m} - (p - r)\frac{1}{\lambda_m}e^{-p/\lambda_m} \tag{1.40}$$

$$0 = 1 - (p - r)\frac{1}{\lambda_m} \tag{1.41}$$

$$(p - r) = \lambda_m \tag{1.42}$$

Intuitively, firms tend to choose higher markups when the unit cost of information processing λ_m is high. When $\lambda = 0$ consumers pay full attention to all alternatives in the market and the outcome is Bertrand competition. The many firms that are tied for the lowest cost capture the entire the market.

Chapter 2

Consumer Search and Switching Costs in Equilibrium

Consider a market in which consumers engage in simultaneous search for a single unit of a differentiated product. Each firm has one product for sale and takes the nonprice characteristics of this product as fixed. Each consumer begins with unbiased estimates of (1) the nonprice value to the consumer of each product and (2) the price the consumer would be charged for each product. Searching a product reveals an idiosyncratic consumer-firm match shock: either a cost shock that is passed through to the consumer or a direct shock to consumer utility.

This paper shows how equilibrium outcomes in such a market depend on consumers' willingness to search for competing offers and to switch their purchases to a competing firm, with an empirical application to the auto insurance industry. I extend a search and switching cost demand model estimated by Honka (2014) with the addition of a supply side, allowing auto insurers to respond in equilibrium to changes in consumer search and switching costs.

The analyis proceeds in two steps. First, I use the estimated demand parameters of Honka (2014) to estimate each firm's marginal cost of selling an additional insurance policy. I assume that firms maximize their present discounted profits and present a method for calculating market shares in future time periods, based on the insight that consumers in the Honka (2014) model transition from firm to firm following a Markov process. I present a simulation-based method for evaluating the transition matrix, from which the entire history of market shares can be easily recovered by matrix multiplication.

With demand and cost estimates in hand, I evaluate the equilibrium impact of counterfactual changes in search and switching costs. In equilibrium, each of the fourteen insurers modeled chooses a mean price offer to maximize its present discounted profit given the

mean price offers chosen by other firms. To efficiently solve for the equilibrium, I take advantage of a typical property of most common demand models: namely, that computing elasticities imposes little or no additional burden when computing market shares. This property provides the basis for an efficient solution process I label *approximate best response iteration*—in fact the multivariate version of Newton's method. Papers proposing this method for solving games include Li and Başar (1987), Baldick and Hogan (2004), Facchinei, Fischer and Piccialli (2009), while Aguirregabiria and Ho (2012) demonstrate its usefulness in evaluating the equilibrium response of the airline industry to changes in various costs. While I do not determine whether the model has a unique equilibrium, I argue that the equilibrium found by approximate best response iteration starting at the observed prices is the most realistic outcome if firms make discrete price adjustments informed by demand elasticities at their current prices.

A frictionless market is not necessarily a benefit to consumers or a detriment to firms, and I find that lower search or switching costs cause equilibrium prices to increase. In the case of search costs, this effect is due to the role of search in differentiating firms. When consumers have low search costs, firms compete less to be searched and more to be chosen following the consumer's (simultaneous) search. This latter competition is less sharp as the information revealed by search serves to differentiate the firms. A crucial assumption is that consumers are informed about each firm's mean price offer before beginning their search. If consumers are completely uninformed about prices before beginning their search, evidence from the sequential search models of Wolinsky (1986) and Haan et al. (2017) suggests that higher search costs will lead to higher prices.

Equally counterintuitively, I find that auto insurers would choose higher price offers and earn higher profits in equilibrium if switching costs were reduced. Consumer loyalty motives firms to exploit currently loyal customers with high prices, but also to gain new loyal customers with low prices. These goals are in conflict given the reputational and often regulatory obstacles to a dynamic or discriminatory pricing strategies in many industries, including auto insurance. Empirically, I find that the business development effect wins out in equilibrium. Halving the preference for remaining with the same firm would increase the price of six months of coverage by $28, or 5 percent.

The results provide support for voluntary efforts by firms to lower consumer search and switching costs across their industries. Websites that facilitate comparison may increase seller profits in a differentiated marketplace by communicating the particulars of product differentiation to consumers. Industries can also gain slightly higher profits by facilitating consumer switching, although eliminating switching costs is estimated to generate a smaller profit increase than eliminating search costs. Generally, initiatives to facilitate switching

might include product standardization, streamlined application and purchase processes, or data transferability for tracking programs such as Progressive Snapshot or American Family's KnowYourDrive that adjust premia in response to driving behavior.

The remainder of the paper proceeds as follows. Section 2.2 extends the Honka (2014) model to include dynamics and a supply side. Section 2.3 illustrates in a simplified, static version of this model why increasing search costs can either increase prices (by making weaker firms less relevant as competitors) or decrease prices (by concealing from consumers some aspects of product differentiation). Section 2.4 summarizes the data, Section 2.5 presents the estimation strategy, and Section 2.6 presents estimation results including counterfactual simulations.

Section 2.7 relates the results to the theoretical literature on search and switching. I argue that the decreasing relation between market frictions and equilibrium prices is due to the assumed *ex ante* observability of prices (for search costs) and the assumption that firms cannot pursue a "bait and switch" pricing strategy (for switching costs). Section 2.8 concludes with implications for policy and modeling.

2.2 Model

This section defines a two-stage model of directed simultaneous search and product choice, closely following Honka (2014). Each consumer demands one six-month auto insurance policy. Consumer i purchasing at firm j has utility

$$u_{ijt} = \alpha_j - \beta W_{ij,t-1} - \gamma p_{ijt} + \varepsilon_{ijt} \tag{2.1}$$

The indicator $W_{ij,t-1}$ takes value 1 if the consumer switches insurers, i.e. was not insured by firm j in period $t-1$.[1]

Consumers initially know all components of u_{ijt}, except p_{ijt} which they learn by searching. Here p_{ijt} represents the price adjusted for various consumer characteristics likely to be associated with the insurance risk and premium. Examples include the number of drivers on the policy, their history of accidents and traffic tickets, and whether any driver on the policy is under 25 years old.[2] Firm j's adjusted price offer is distributed (independently

[1] Honka (2014) equivalently defines β as the positive addition to utility for not switching. I adopt the opposite convention so that a lower β represents a reduction in search costs, directly making switching consumers better off and nonswitching consumers no worse off.

[2] Details of this adjustment are in Honka (2014).

across firms) as

$$p_{ijt} \sim \text{Gumbel}(\overline{p}_j - \mu e_c, \mu) \tag{2.2}$$

where $\overline{p}_j$ is a baseline price offer chosen by firm j, μ is a scale parameter determining the dispersion of prices within a firm, and $e_c \approx 0.577$ is Euler's constant. The mean of this distribution is $\overline{p}_j$. The remaining terms in Equation 2.1 are a brand value α_j and a preference shock ε_{ijt}. Consumers initially know α_j, ε_{ijt}, and $\overline{p}_j$ but discover p_{ijt} only by searching firm j.

Define firm j's *ex ante* utility to consumer i as

$$u_{ijt}^{\text{ante}} = \alpha_j - \gamma \overline{p}_j + \varepsilon_{ijt} \tag{2.3}$$

.

From the perspective of a consumer choosing a firm to search, the *ex post* utility u_{ijt} is distributed as u_{ijt}^{ante} minus γ times a mean zero price shock. The distribution of u_{ijt} thus differs across firms only by a location shift, allowing firms to be ranked by first order stochastic dominance. Applying the intuitive selection rule of Chade and Smith (2006), consumers direct their search to the firms with highest u_{ijt}^{ante}.

To rationalize the fact that consumers do not search all firms, assume that the consumer's utility is diminished by a search cost c for every firm searched. Following Honka (2014), I make an exception for the consumer's previous insurer and assume that the consumer automatically searches it without paying a search cost (by receiving a renewal notice or by automatically renewing). Overall the consumer's expected utility is

$$\text{E}[u_i] = \text{E}[\max_{j \in K_i} u_{ijt}] - c(|K_i| - 1) \tag{2.4}$$

where K_i is the set of firms searched and $|K_i|$ its size. The optimal K_i can be found by starting with the previous insurer and adding firms in order of decreasing u_{ijt}^{ante}, stopping when an addition would decrease expected utility.

I turn now to the firm's problem. Assume that if firm j serves consumer i, it receives a price of p_{ijt} and incurs a cost of r_{ijt}. At its broadest, the firm's problem could involve choosing a price p_{ijt} for each consumer and time period. Instead I take as fixed the variation in a firm's price offers and consider only each firm's choice of its mean price offer $\overline{p}_j$. Formally, let $\overline{r}_j$ be a base level of marginal cost for each firm and assume

Assumption A1. $p_{ijt} - r_{ijt} = \overline{p}_j - \overline{r}_j$

Assumption A1 equivalently states that the price shock $(p_{ijt} - \overline{p}_j)$ represents a full pass-through of additional costs incurred by the firm from the match between consumer i and firm j. These costs may be real or merely perceived. Insurers for instance may differ in the amount and type of coverage offered by their standard plans. This could make certain consumers more costly for some insurers than for others. Insurers may also differ in their risk assessments of a given consumer and may pass on these perceived risk differences to their quoted premiums.

The benefit of Assumption A1 is to simplify the firm's profit maximization problem in two ways. First, all consumers are equally profitable to serve. Firms can thus consider only their overall market shares rather than considering which particular consumers are attracted by a particular pricing policy. Second, the firm's problem of setting many prices p_{ijt} is reduced to the choice of a single baseline price offer $\overline{p}_j$.

An opposite assumption would be that variation in $(p_{ijt} - \overline{p}_j)$ represents price discrimination unrelated to the cost or risk of providing the service. Honka (2014) notes for instance that some auto insurers offer lower premiums to new customers. However, state regulations commonly require that premium differences be justified on the basis of cost, which may limit the scope of price discrimination. The possible presence of price discrimination in auto insurance is discussed further in Section 2.7.

Letting $\overline{p}$ be the vector of all firms' prices, firm j maximizes its discounted stream of profits for the next T six-month periods.

$$\max_{\overline{p}_j} \sum_{t=1}^{T} \delta^t \sum_i \Pr(i \text{ chooses } j|\overline{p})(p_{ijt} - r_{ijt}) \tag{2.5}$$

Equation 2.5 can be written using Assumption A1 as

$$\max_{\overline{p}_j} (\overline{p}_j - \overline{r}_j) \sum_{t=1}^{T} \delta^t \sum_i \Pr(i \text{ chooses } j|\overline{p}) \tag{2.6}$$

I set $\delta = \frac{1}{\sqrt{1.06}}$ for an annual discount rate of 6 percent and $T = 100$ for a 50-year driving career of an individual consumer. While these parameter choices are somewhat arbitrary, the data rejects the assumption that firms maximize only their immediate profits. Setting $T = 1$ in the estimation (Section 2.5) produces unreasonably low marginal cost estimates for the major insurers. For the largest insurer Geico, assuming infinite discounting ($T = 1$) leads to a negative estimate of marginal cost. Cost estimates under alternative assumptions on discounting are summarized in Table 2.8.4.

I assume that each consumer draws new preferences in each period. Any persistence in

consumer behavior is captured in reduced form by the inertia value of choosing the same insurer as in the previous period. Consumer movements across firms thus form a Markov process with transition matrix P. The vector of market shares in time period t is $s_t = s_0 P^t$, allowing for an efficient evaluation of a firm's objective function in 2.5

The model applies equally to settings in which search reveals an additional nonprice preference shock rather than a price shock resulting from a cost shock. For instance, a car buyer may search by taking test drives that reveal each vehicle's comfort and handling (Moraga-González et al., 2015).

2.3 More search may increase or decrease prices

Consumer search in this model serves to reveal a price shock representing a passed-through cost shock, which from the firm's perspective is equivalent to informing the consumer about additional product differentiation. I argue that consumer search of this sort has an ambiguous effect on equilibrium prices, justifying this paper's empirical investigation of the effects.

A simple static model provides intuition for this ambiguity. Consider a market with two firms and no outside option. Each firm j has zero marginal cost and offers a product of quality q_j, with $q_1 \geq q_2 = 0$. Suppose that if consumers know only the price and quality of each product, they choose according to the utility function $u_{ij} = q_j - p_j$. If $q_1 > q_2$ then the standard Bertrand-Nash result is an equilibrium price p_1^* slightly less than q_1, with Firm 1 capturing the entire market. If $q_1 = q_2$ then $p_1^* = p_2^* = 0$.

Now suppose that instead of choosing based on price and quality alone, consumers search both firms and discover some aspect of horizontal product differentiation. In particular, consider the case of the Hotelling (1929) model and suppose that each product has a location or variety that has been discovered by search. Assume that Firm 1 is at location 0, Firm 2 is at location 1, and that consumers are uniformly distributed along the unit interval with travel cost equal to the distance to the selected firm. It is well known that product differentiation of this sort, if revealed to consumers, can increase equilibrium prices. If both firms have quality $q_1 = q_2 = 0$ for instance, the equilibrium price will increase from $p_1^* = p_2^* = 0$ in the undifferentiated Bertrand case to 1 in the Hotelling model.

On the other hand the Hotelling model may produce lower prices than the Bertrand model, provided that one of the firms is much weaker than the other. Suppose $q_1 = 2$. If in the Hotelling model Firm 1 attempted to charge its undifferentiated Bertrand equilibrium price of 2, a consumer near location 1 would have utility slightly above -1 from purchasing from Firm 1. An opening is created for Firm 2 to gain positive market share, unless Firm 1

reacts.

By a standard derivation for the Hotelling model, the equilibrium prices as a function of q_1 are $p_1^* = 1 + \frac{q_1}{3}$, $p_2^* = 1 - \frac{q_1}{3}$. This implies $p_1^* = \frac{5}{3}$ for $q_1 = 2$, which is less than the Bertrand price. The Hotelling equilibrium still has all consumers choosing Firm 1 as in the Bertrand case. Yet by creating a group of customers especially interested in Firm 2, the Hotelling model improves Firm 2's competitive position and forces Firm 1 to lower its price.

A notable threshold in this comparison is at $q_1 = \frac{3}{2}$. For $q_1 < \frac{3}{2}$, the Hotelling model produces a higher equilibrium price than the Bertrand model. Competitors of similar quality levels raise their prices in the presence of search due to a well known *product differentiation effect*. When $q_1 > \frac{3}{2}$, a *relevant competitors effect* dominates. Informing consumers about horizontal differentiation randomly perturbs the market, to the advantage of firms whose products were unpopular before the perturbation.[3]

2.4 Data

This section summarizes the data and demand estimates of Honka (2014), which are used as inputs to the full equilibrium model of this paper. Honka uses detailed survey data on consumer search and purchases to estimate both search costs and the cost of switching to a new insurance provider.

For each consumer, Honka (2014) observes the current insurer, previous insurer, and a list of insurers from whom the consumer obtained quotes. Premiums are observed only for the insurer the consumer chooses, although Honka reconstructs likely premiums quoted by the other firms. Premiums are adjusted for consumer characteristics such as number of past accidents, living in an urban area, and being under 25 years old. Combining the data on search, choice, and price, Honka estimates several demand models. I apply Honka's simplest demand model "Model 0", described in the previous section). Model 1 allows for arbitrary correlations in consumer tastes ε_{ijt} for different insurers, while Model 2 additionally relates consumer preferences to demographic and psychographic differences such as the consumer's interest in finance and satisfaction with the previous insurer.

As in Honka (2014), I assume consumers search simultaneously rather than sequentially. Honka and Chintagunta (2017) provide evidence that simultaneous search is a more plausible data generating process for the Honka data. Sequential search would imply that consumers who stopped searching after the first search were more likely to have a below-

[3] Economides (1989) considers a Hotelling model in which firms choose differentiated locations, quality levels, and prices. In equilibrium, firms choose as much differentiation as possible.

average price draw conditional on their choice of firm. The authors find that this pattern does not appear in the data.

In order to model insurance competition at the national level, this paper excludes three firms noted by Honka (2014) as having a limited geographic range of operation: American Family, Erie, and Mercury. The combined market share of these firms was 8 percent. Table 2.8.2 reports basic characteristics of the insurers: their average price offers, market shares, and brand values (including any effects attributed to advertising in Honka (2014)). Table 2.8.3 reports other parameters of the demand model such as the estimated search and switching costs.

2.5 Estimation

Taking as given the estimated demand "Model 0" of Honka (2014) (with the reduced switching cost noted in the previous section), I estimate each firm's base level of marginal cost $\overline{r}_j$ corresponding to a consumer with $p_{ijt} - \overline{p}_j = 0$. The cost estimates are used later to model the equilibrium response of firms to changes in search or switching costs.

Estimation begins by finding the Markov transition matrix P of consumer movements across firms. The element $P_{kj'}$ is the market share of firm j' among consumers who were insured by firm k in the previous period. As noted by Honka (2014) there is no obvious analytic solution for these market shares, which necessitates the use of Monte Carlo simulation for both taste shocks $\varepsilon_{ij't}$ and price shocks $(p_{ij't} - \overline{p}'_j)$. Honka (2014) uses a kernel smoothed frequency simulator for maximum likelihood estimation of the demand model. However, cost estimation requires computing not only the market shares but their derivatives with respect to prices, which motivates the different simulation approach described below.

To find $P_{kj'}$, I take simulation draws of the ε_{ijt} of all firms other than j', and for the $(p_{ijt} - \overline{p}_j)$ of all firms. Holding out $\varepsilon_{ij'}$ from simulation retains some randomness in the purchase decision, leading to a non-degenerate market share of firm j' and allowing for the evaluation of derivatives. Fixing a single simulated consumer i, the calculation proceeds in two steps.

First, identify for every value of $\varepsilon_{ij'}$ the optimal set of firms to search. The Chade and Smith (2006) rule of selecting the top-ranked firms simplifies this problem. For the lowest values of $\varepsilon_{ij'}$, the optimal search set is the set that would be searched if firm $\varepsilon_{ij'}$ did not exist. This set is found by starting with the initial firm and adding firms from (*ex ante*) best to worst, until expected utility no longer increases. As $\varepsilon_{ij'}$ increases, firm j' enters the search set either as an addition or as a replacement for the worst firm in the search set (in

terms of *ex ante* utility).[4]

As $\varepsilon_{ij'}$ grows further, it becomes optimal to successively drop the *ex ante* worst firms in the search set. Intuitively, searching a firm other than j' becomes less valuable as j' becomes more likely to be the *ex post* best firm. The worst firms are dropped from the search set until only firm j' remains—along with the previous insurer, which is automatically searched. I solve numerically for the thresholds of $\varepsilon_{ij'}$ at which the optimal search set changes. Table 2.8.1 shows one simulated consumer's possible optimal search sets, as well as the thresholds of $\varepsilon_{ij'}$ that separate these search sets.

Second, I find the probability of accepting the offer of firm j' conditional on searching a given set of firms K_i. Again taking as given ε_{ijt} for all $j \neq j'$, I iterate over vectors of simulated price shocks $(p_{ijt} - \overline{p})$ and find the average conditional probability that firm j has the highest utility. As discussed above (and illustrated in Table 2.8.1), the decision to search set K_i implies that $\varepsilon_{ij'}$ is within some known range. It is this truncated distribution of $\varepsilon_{ij'}$ that is used to compute the following conditional acceptance probability.

$$\begin{aligned} &\Pr(j' \text{ accepted} | K_i \text{ searched}, \varepsilon_{ijt}\ \forall j \neq j', W_k = 0) \\ &= \textstyle\sum_q Pr(j' \text{ accepted} | K_i \text{ searched}, \varepsilon_{ijt}\ \forall j \neq j', (p_{ijt} - \overline{p}) = q, W_k = 0) \end{aligned} \tag{2.7}$$

The final probability of simulated consumer i searching and then purchasing from firm j' is

$$\begin{aligned} &\Pr(j' \text{ accepted} | \varepsilon_{ijt}\ \forall j \neq j', W_k = 0) \\ &= \textstyle\sum_{K_i} [\Pr(K_i \text{ searched} | \varepsilon_{ijt}\ \forall j \neq j', W_k = 0) * \\ &\Pr(j' \text{ accepted} | K_i \text{ searched}, \varepsilon_{ijt}\ \forall j \neq j', W_k = 0)] \end{aligned} \tag{2.8}$$

Finally, consumer-specific market shares are averaged over the simulated consumers to give $P_{kj'}$. The derivative $\frac{dP_{kj'}}{d\overline{p}_j}$ is found similarly by averaging over simulated consumers, with the consumer-specific price derivative found by differentiating 2.8 using the product rule. Details of the price derivatives of the search and conditional acceptance probabilities are in Appendix 2.A.

The simulation uses 200 Gumbel$(0, 1)$ vectors of taste shocks ε_{ijt} and 1,000 Gumbel$(\overline{p}_j - \mu e_c, \mu)$ vectors of price shocks $(p_{ijt} - \overline{p}_j)$.[5] After computing the transition matrix P of mar-

[4] I find the value of $\varepsilon_{ij'}$ at which the consumer would be indifferent about replacing the worst searched firm with j'. If at this value of $\varepsilon_{ij'}$ the consumer prefers to add j' to the search set without dropping the worst firm, then firm j' enters as an addition rather than a replacement (and may enter at a different $\varepsilon_{ij'}$ value, which I again find numerically).

[5] Simulation draws were taken from a Halton sequence to reduce sampling error.

ket shares conditional on the previous insurer, I find each firm's present discounted profit as defined in Equation 2.5. The vector of market shares at time t is given by

$$\boldsymbol{s}^t = P^t \boldsymbol{s}^0 \tag{2.9}$$

where $\boldsymbol{s}^0 = P^{-1}\boldsymbol{s}^1$ and $\boldsymbol{s}^1$ is the vector of observed market shares.

Given P and its own and cross-price derivatives $\frac{dP_{jk}}{dp_m}$, I estimate each firm's marginal cost $\overline{r}_j$ of insuring a typical consumer. Define

$$q_j = \sum_{t=0}^{T} \delta^t \sum_i \boldsymbol{s}_j^t \tag{2.10}$$

as the present discounted market share stream of firm j. Equation 2.6 can then be written as

$$\max_{\overline{p}_j} (\overline{p}_j - \overline{r}_j) q_j \tag{2.11}$$

implying the standard first order condition

$$0 = q_j + (\overline{p}_j - \overline{r}_j) \frac{dq_j}{d\overline{p}_j} \tag{2.12}$$

.

The value of $\frac{dq_j}{d\overline{p}_j}$ is found by automatic differentiation of Equation 2.10, using the expression for market shares in Equation 2.9.[6]. In addition to estimating costs, Equation 2.12 provides the basis for this paper's counterfactual predictions of how firms would change their price offers given a change in consumer search or switching costs.

2.5.1 Approximate best response iteration

Consider the supply-side response to a counterfactual change in the parameters of the demand model. At the new market equilibrium, prices $\overline{p}$ are such that equation 2.12 is satisfied for every firm j. To find $\overline{p}$ satisfying these conditions, I use an approximate version of best response iteration that takes a first-order Taylor approximation to $\frac{dq_j}{d\overline{p}_j}$. This approach was shown by Baldick and Hogan (2004) to either find a Nash equilibrium (to within the user's choice of precision) or fail to converge.

Each iteration begins with a current vector of prices $\tilde{p}$, at which I evaluate discounted

[6] Automatic differentiation is by the `ForwardDiff` package in Julia.

market shares $\tilde{q}$ and $\frac{dq}{d\overline{p}}|_{\tilde{p}}$. Recall the firm's optimization problem 2.11 and apply a Taylor approximation around $\tilde{p}$.

$$\max_{\overline{p}_j}(\overline{p}_j - \overline{r}_j)(\tilde{q}_j + (\overline{p}_j - \tilde{p}_j)\frac{dq_j}{d\overline{p}_j}|_{\tilde{p}}) \tag{2.13}$$

Taking the first order condition gives a closed form solution for the price $\overline{p}_j$.

$$0 = (\tilde{q}_j + (\overline{p}_j - \tilde{p}_j)\frac{dq_j}{d\overline{p}_j}|_{\tilde{p}}) + (\overline{p}_j - \overline{r}_j)\frac{dq_j}{d\overline{p}_j}|_{\tilde{p}}) \tag{2.14}$$

$$0 = 2\overline{p}_j\frac{dq_j}{d\overline{p}_j}|_{\tilde{p}} + \tilde{q}_j - (\tilde{p}_j + \overline{r}_j)\frac{dq_j}{d\overline{p}_j}|_{\tilde{p}} \tag{2.15}$$

$$\overline{p}_j = \frac{1}{2}((\tilde{p}_j + \overline{r}_j) + \frac{\tilde{q}_j}{\frac{dq_j}{d\overline{p}_j}|_{\tilde{p}}}) \tag{2.16}$$

Approximate best response iteration updates all prices simultaneously by this method and uses the solution as the next candidate $\tilde{p}$. This process continues until convergence, requiring only one evaluation of the transition matrix P and its derivatives on each iteration. Toward the equilibrium, the difference between the initial candidate $\tilde{p}$ and the new solution p approaches zero. This implies that the error introduced by the Taylor approximation also approaches zero. For all counterfactuals, convergence was achieved in 8 iterations or fewer.[7]

Approximate best response iteration can be used to solve any sufficiently well-behaved system of nonlinear equations, where it is more generally known as the multivariate Newton method. However, it is particularly suitable to finding Nash equilibria in economic models, as demonstrated by Aguirregabiria and Ho (2012) in solving a game of airline competition. Each airline is assumed to have a local manager for each city pair, whose (mixed) strategy is the probability with which to operate flights between the cities. Despite the very large number of players in the game (32,670), approximate best response iteration succeeds in identifying an equilibrium.

For typical models in industrial organization, approximate best response iteration has several advantages over competing methods. Computing market shares and profits is often computationally costly, preventing the explicit computation of best responses needed by the Jacobi and Gauss-Seidel methods. Yet once profits are calculated, the own-price derivatives of the profit functions tend to be readily available as these are also needed for cost estimation. In this paper's problem, evaluating P may take up to 15 minutes, but also

[7]Convergence was defined as no price changing by more than \$1 between iterations.

finding $\frac{dP}{dp}$ does not significantly increase the computational burden.[8] In all counterfactuals, convergence was achieved in 8 iterations or fewer.

In addition, approximate best response iteration incorporates an intuitively convincing equilibrium selection rule. Given a counterfactual change in parameters, the selected equilibrium is that which is reached by firms starting from the observed prices and iteratively maximizing a heuristic profit function based on currently observed elasticities. In the case that approximate best response iteration fails to converge, the researcher has recourse to more complex derivative-based methods such as the global Newton method (Govindan and Wilson, 2003) implemented in specialized software such as Gambit.

2.6 Results

Table 2.8.5 presents observed market shares and prices alongside estimates of present discounted market shares and marginal costs, taking as given the demand model and estimates of Honka (2014). In some cases the discounted market share is quite different from the observed market share. For instance, the model predicts that relatively small insurer 21st Century will surpass the market leader Geico if both maintain their pricing policies and brand values.

The cost estimates in Table 2.8.5 are used as inputs to counterfactual simulations of different search and switching costs. Honka (2014) estimates a search cost of $c = 0.1663$ in terms of utility, equivalent to \$42 per firm searched. The switching cost is $\beta = 1.3375$, or \$336. These sizable frictions help to explain the 74 percent retention rate found by Honka (2014). Suppose counterfactually that search costs are reduced, for example by the increased availability and use of online shopping since the 2006 data collection of Honka (2014), or by the rise since the early 2000s of price comparison websites. The rise of online shopping may have also decreased switching costs.

Table 2.8.7 presents the results of the counterfactual simulations. Comparing Cases 1 and 2 with the baseline equilibrium shows that prices would be increased by reducing or removing switching costs. Intuitively, firms compete more fiercely for loyal consumers than for consumers with little or no inertia. Moreover, this pro-competitive effect outweighs the benefit to firms of attempting to exploit their currently loyal customers by setting a high $\overline{p}_j$. This intuition is discussed further in Section 2.7, along with its connection to the literature on switching costs.

Cases 3 through 9 of Table 2.8.7 explore the impact of changing search costs (while maintaining the estimated switching cost). Perhaps counterintuitively, firms increase their

[8] Differentiation of P is discussed in Appendix 2.A.

prices as search costs decrease. The explanation lies in the information available to consumers. At higher search costs, consumers tend to narrow their options more at the search stage before the $(p_{ijt} - \overline{p}_j)$ are known. Firms are thus more differentiated at the choice stage than at the search stage, implying that firms charge higher markups when they compete primarily at the choice stage. Section 2.7 returns to this argument in detail.

The last counterfactual (Case 10) simulates a frictionless market with no search or switching costs (although the initial firm still offers a free search through a renewal offer). Curiously the mean price offer is lower than if only switching costs are removed (Case 3), suggesting an anti-competitive effect of search costs in this case. One explanation is that without inertia to retain their current customers, some firms are likely to be uncompetitive due to high costs or poor quality. A low or zero search cost may be necessary for these weak firms to be searched and to offer real competition to firms with low costs and high quality products.

The consumer impact of removing search and switching costs depends on how switching costs are interpreted. One view is that these costs consist mainly of real money and time costs, and that a rational consumer would indeed pay $336 as Honka (2014) estimates to avoid switching. This view seems at odds with the fact that switching can be done in minutes (or at most, a few hours) once the consumer has searched the firm. Woodward and Hall (2012) make a similar argument after estimating that mortgage consumers sacrifice over $1,000 by shopping from too few mortgage brokers. Following Woodward and Hall (2012), implied consumer "costs" that are very disproportionate to the time and expense involved may be better interpreted as resulting from behavioral biases or from unawareness of the opportunities available.

Column 4 of Table 2.8.7 reports consumer surplus if search and switching costs are interpreted as real costs, while Column 5 excludes them as not representative of an informed and rational consumer's welfare. If search and switching costs are real costs, removing switching costs (Case 2) benefits consumers in the amount of $111 as the ability to switch freely outweighs the impact of increasing equilibrium prices. If instead search and switching costs are viewed as behavioral only, consumer surplus is essentially unaffected by the removal of switching costs.

Removing search costs would benefit consumers despite increasing prices. Comparing the baseline case to the removal of search costs in Case 9, consumers gain surplus of $22 including search and switching costs or $50 excluding these costs. However, merely reducing search costs may result in lower consumer welfare, as shown by comparing the baseline case to Case 6. The consumer harm from higher prices may not be outweighed by the newly increased tendency to search, or by the (real or imagined) benefit of paying

less in search costs. In Case 10, removing all frictions benefits consumers relative to the baseline.

Some firms are more responsive than others to changes in search and switching costs. Firms can be classified as strong or weak according to the total surplus created by a transaction between the firm and a baseline consumer with $\varepsilon_{ijt} = 0, (p_{ijt} - \overline{p}_{jt}) = 0$. The strength of firm j is the total surplus

$$\alpha_j - \gamma \overline{r}_j \tag{2.17}$$

Figure 2.8.1 shows that "strong" firms whose transactions create high total surplus tend to increase their prices more than other firms when switching costs are removed from the market, but perhaps slightly less than other firms when search costs are removed. One interpretation is that switching costs advantage weak firms, who may benefit from a single advantageous taste shock ε_{ijt} in one time period and retain the consumer beyond that period despite generally offering lower quality and higher prices. With switching costs removed, weak firms become less relevant as competitors, allowing strong firms to increase their prices.

2.6.1 Applicability of the estimates

The firms I identify as "strong" with high total surplus do not necessarily have high market shares in Honka's 2006-2007 data. Geico for instance is far from the strongest firm but has the highest market share. The subsequent evolution of market shares also fails to support the classification of firms by strength. In addition to total surplus, Table 2.8.5 shows each insurer's share in the 2006-2007 data and (for the largest insurers in 2018) its 2018 market share. State Farm was predicted to lose most of its market share, but instead became the market leader by 2018. Some particularly strong firms such as The Hartford and Liberty Mutual were expected to more than double their market shares but instead saw their market shares decline.

Changes in costs, strategies, and consumer tastes may partly explain the lack of convergence of market shares to their predicted values. Additionally, the large insurers may have benefitted from consumers newly entering the market with only limited awareness of the brands available. Newly entering consumers are not included in the survey data. Another explanation is that some model parameters are incorrectly estimated. For instance, a firm that discounted its future earnings sharply and strategically chose high markups would have its marginal cost overestimated by this paper's method. Regardless of the validity of

the 2006-2007 estimates, it is clear that the market has evolved substantially since then.[9] Counterfactual simulations based on the estimates are of limited value for assessing potential impacts on today's insurance market. Nonetheless these simulations offer qualitative insights into the relation between search and switching costs and market outcomes.

Another way to gauge the realism of the results is to compare marginal cost estimates to costs reported by firms in their annual reports. For example, Progressive in 2006 reported claims expenses (losses and loss adjustments) equal to 66.5 percent of revenue, plus underwriting expenses equal to an additional 20.1 percent of revenue. The marginal cost of writing an additional policy to a typical customer should thus equal 66.5 percent of the premium to pay expected claims, plus between zero and 19 percent of the premium to cover additional underwriting resources needed to originate and service the policy.

I find that the (sales-weighted) average of marginal costs is 65 percent of the average premium paid, a value which Table 2.8.6 suggests may be too low. Annual reports of the four largest insurers from around the time of Honka's data collection show average claims expenses of around 70 percent of premiums and underwriting expenses of around 20 percent of premiums, suggesting that the average marginal cost should lie between 70 and 90 percent of the average premium paid. The model fails to explain why insurers do not choose higher prices in light of the demand estimates, with estimation resorting to somewhat low marginal cost estimates as an explanation. Accurate cost estimation and accurate predictions may require starting with a significantly more detailed demand model such as Honka's "Model 2" with persistent consumer heterogeneity. The approach I propose in this paper generalizes naturally to a setting with multiple consumer types, by constructing a different Markov transition matrix for each type. Incorporating additional types will naturally increase the computational and data requirements, making the continuous type space of Model 2 intractable under this approach unless the type space is reduced. A question raised in Honka (2014) is how to interpret the large inertia term β, estimated at 1.3375 in Model 0 and equivalent to a price increase of $336 on a six-month policy. Honka's "Model 2" models the switching cost as a linear combination of several demographic and psychographic measures, as well as the consumer's satisfaction with various aspects of the insurer's service. Honka attempts to separate inertia β into switching costs and satisfaction-based inertia, concluding that only about 10 percent of inertia is attributable to switching costs. However, all that can properly be concluded from such an exercise is the difference in inertia between customers of different satisfaction levels. A truly dissatisfied customer will be more likely to switch firms, and might indeed exhibit as little as 10 percent as much

[9] Technological progress in auto insurance must also be considered. For instance, some insurers now offer in-car trackers that allow insurance rates to vary based on driving habits.

inertia as a more satisified consumer as Honka (2014) finds. Yet the full inertia of a typical customer can rightly be labeled a switching cost. This inertia is found in Honka's Model 0 when all consumers are assigned the same inertia. Efforts to attribute part of this typical consumer's inertia to customer satisfaction must be limited to relative comparisons such as how much inertia is reduced by the typical customer being less than fully satisfied.

2.7 Comparison with other models of search and switching

2.7.1 Search costs

The results above can be explained in terms of two largely separate theoretical literatures on differentiated product market equilibrium, in the presence of either search costs or switching costs. Within the search cost literature, the closest parallel to this paper (and to Honka (2014)) is Haan, Moraga-González and Petrikaite (2017). The main differences are that Haan et al. (2017) assume (1) a symmetric duopoly model rather than this paper's 14 firms with different α_j and r_j and (2) sequential rather than simultaneous search. Haan et al. (2017) establish theory and intuition for my result that higher search costs may tend to lower prices and profits. Further, they explain how this result depends on what information is available to consumers before they begin ther search. Haan et al. (2017) is one of many papers extending Wolinsky (1986), in which consumers search to learn both prices and idiosyncratic nonprice match values. In the Wolinsky model, consumers initially know nothing about the individual firms and engage in random or undirected search. As observed by Anderson and Renault (1999), the Wolinsky model "yields intuitive comparative statics results: the equilibrium price rises with search costs and falls with the number of firms". Unlike in the Diamond (1971) model for homogeneous products, a small search cost generates only a small increase in prices. The contribution of Haan et al. (2017) is to show that if consumers are informed about prices before searching (and can direct their search accordingly by the Weitzman (1979) rule), prices instead fall as search costs increase.[10]

Haan et al. (2017) offer intuition for this reversal based on differences in the strength of consumer preferences for ex ante observable nonprice quality (ε_{ijt} in this paper's notation). To summarize and reframe this intuition, increased consumer search creates a perception of product differentiation. If search costs are so high that consumers choose to search only once, they effectively commit to a single firm at the search stage before learning any later-

[10]There are certain conditions on this result. Either the density of pre-search observed match values must be log-concave or the search cost must be high enough.

realized shocks: in this paper, $(p_{ijt} - \overline{p}_j)$. Because I assume $p_{ijt} - \overline{p}_j$ is a passthrough of a cost shock $r_{ijt} - \overline{r}_j$, it does not directly affect the firm's profit and has the same effect on the model as the post-search taste shock assumed by Haan et al. (2017).

Whenever consumers must choose without observing the full extent of product differentiation, market power is eroded and equilibrium prices are driven down. Haan et al. (2017) reviews several empirical papers that estimate an increase in search costs would either decrease prices (Pires (2018), Moraga-González et al. (2017)) or increase price elasticities (Dubois and Perrone (2018), Koulayev (2014)). Papers in which search is for prices and match values, or prices alone, are more likely to find an increasing relationship between search cost and price. Allen, Clark and Houde (2014) and Alexandrov and Koulayev (2017) for instance assume that mortgage borrowers learn interest rates through search and find that lower search costs would reduce equilibrium interest rates. For researchers and policymakers, these contrasting results suggest the value of survey or other evidence on the search process and on which product attributes consumers seek to learn by searching. Alexandrov and Koulayev (2017) for instance (and Chapter 1 of this dissertation) take advantage of a survey of mortgage borrowers in estimating search models. Likewise for firms and marketers, strategies of advertising and transparency should consider the sort of information being revealed. While firms may benefit from obfuscating their overall level of prices (as in Ellison and Wolitzky (2012)), they will face increased competition and reduced profits to the extent this obfuscation also hinders consumers from learning the horizontally differentiating characteristics of each firm.

2.7.2 Switching costs

A separate literature originating with von Weizsäcker (1984) and Klemperer (1987) addresses the cost to a consumer of switching firms in a dynamic model. Klemperer (1987) observes that switching costs weaken competition as the market becomes segmented into consumers loyal to each firm. However, there is fierce competition in the first period as firms seek to attract consumers who will subsequently become loyal to the firm. In a more general many-period model with the possibility of price discrimination, a firm could offer low "teaser rates" to new customers, to be increased later once the customer is attached to the firm by a switching cost. In contrast, von Weizsäcker (1984) proposes a model in which firms commit to a single price for all time:

> a supplier who establishes a reputation for fair, nonexploitative treatment of customers with switching costs, may be able to overcome the reluctance of users to incur an ongoing cooperative relation with him, if users see that there

> is long run competition between this supplier and other suppliers and that therefore it would be shortsighted for a given supplier to exploit his customers and thereby ruin his reputation.

The auto insurance industry combines some of the inertia exploitation of the Klemperer model with the static fairness of the Weizsäcker model. Insurers are regulated in their rate setting by state insurance commissioners, who may require them to justify departures from a rate schedule. However, these regulations are not uniform in either their requirements or their application. In 2020, the Maryland Insurance Administration rejected Allstate's request to revise its auto insurance premia using a "customer retention" model. An analysis by Consumer Reports and The Markup argues that the Allstate model would have based premium increases on willingness to pay rather than solely on costs. Yet the analysis observed that Allstate already used some form of "customer retention" pricing model in at least ten other states (Varner and Sankin, 2020). What would be called "price discrimination" in standard economic terms has come to be classified under "price optimization" in the language of firms and regulators. Price optimization is however broader and may include using novel consumer-level databases to more accurately predict each customer's claims expenses. This usage is not price discrimination, but rather cost passthrough. Even true price discrimination may be of the static sort that varies prices across customers, rather than of the inertia exploitation sort that increases prices for the same customer at the time of policy renewal.[11]

Given that my source data from Honka (2014) cannot distinguish premia for new policies from premia for renewals, I follow von Weizsäcker (1984) in reducing the firm's pricing problem to a one-shot decision. I obtain von Weizsäcker's counterintuitive result that more consumer inertia leads to more competition, here in the form of lower prices. To the extent that firms do exploit consumer inertia by charging different premia for new policies and renewals, consumer inertia would likely reduce consumer welfare as proposed by Klemperer (1987).

[11] A 2015 report by the National Association of Insurance Commissioners (NAIC) found that 45 percent of large insurers used "price optimization" and a further 29 percent did not use it but had plans to start (National Association of Insurance Commissioners: Casualty Actuarial and Statistical Task Force, 2015). The report also refers to a 2013 survey of major insurers finding that 55 percent considered customer price elasticity when choosing prices. It is unclear whether this question was understood to refer to the overall elasticity of market demand or to the elasticity of demand for specific groups of consumers. The NAIC report recommended interpreting statutory bans on "unfairly discriminatory" prices to include rates based on consumer differences in price elasticity, propensity to shop for insurance, predicted retention rate, or propensity to ask questions or file complaints. Some states such as California, Ohio, and Maryland have issued regulatory guidance banning these practices.

2.8 Conclusion

This paper provides an empirical illustration of the theoretical results on search and switching costs in Haan et al. (2017) and von Weizsäcker (1984). I show that higher search costs still tend to reduce prices, even when there are 14 non-identical firms rather than the symmetric duopoly in Haan et al. (2017). Intuitively, high search costs require firms to compete more to be searched and less to be chosen conditional on being searched. The former contest is inherently more competitive than the latter when consumers use search to learn about product differentiation. In the competition to be searched, firms do not possess the market power that comes from a favorable idiosyncratic draw of a taste shock or fully passed through cost shock. Switching costs also tend to decrease prices as firms compete more intensely to gain loyal customers. Moreover, search and switching costs each continue to exert these equilibrium effects in the presence of the other. I discuss ways in which these results might be altered or reversed: for instance, in the original Wolinsky (1986) model where consumers are uninformed about prices. The pivotal nature of such information assumptions argues for further research into what information consumers know at the outset, what they learn by search, and what they learn only after purchase.

For a firm, this paper's results underline the importance of informing consumers about product differentiation. Although some obfuscation or reticence about pricing may be advantageous, differentiating characteristics of the product should be communicated as efficiently as possible. Norms in certain industries may assist in communicating product differentiation: a common search process or the provision of data to aggregator sites, collocation of firms to ease consumer search (as documented by Murry and Zhou (2016) for auto dealers), and sponsorship of comparison sites and independent reviewers. Ironically, the standardization of certain product attributes may also increase the amount of product differentiation communicated to consumers, as buyers of standardized products are able to search efficiently and focus attention on attributes where meaningful differentiation exists.

This paper's second conclusion is methodological. I develop methods to complete the demand model of Honka (2014) with a supply side, creating a general model of search and switching costs that can be applied to a variety of industries. These methods include an efficient means of computing market shares and their derivatives at a given price (the simulation method of holding out $\varepsilon_{ij't}$), an efficient means of predicting future market shares (the Markov transition matrix), and an efficient means of solving the pricing game (approximate best response iteration). Research that deals only with search costs or only with switching costs may benefit from adopting some subset of these methods.

Search set	Min. $\varepsilon_{ij't}$	Max. $\varepsilon_{ij't}$
$\{1, 10\}$	$-\infty$	1.8055
$\{2, 1, 10\}$	1.8055	3.0164
$\{2, 1\}$	3.0164	$+\infty$

Table 2.8.1: Illustration of optimal search sets, listing the sets that include firm $j' = 2$ for one simulated consumer previously insured by fim 1. The search set is a function of $\varepsilon_{ij't}$.

	α_j	$\overline{p}_j$	Observed share (pct.)
21st Century	3.06	6.31	0.03
AIG	3.34	6.71	0.04
Allstate	3.20	7.07	0.15
Farmers	3.28	6.23	0.05
Geico	2.66	5.89	0.20
GMAC	2.89	7.34	0.02
The Hartford	3.30	6.01	0.05
Liberty Mutual	3.56	7.05	0.05
Metlife	2.78	6.90	0.02
Nationwide	2.72	6.66	0.04
Progressive	3.17	6.47	0.13
Safeco	2.38	6.42	0.02
State Farm	2.91	7.04	0.14
Travelers	3.19	7.28	0.05

Table 2.8.2: Results from Model 0 of Honka (2014): nonprice quality α_j, mean price offers, and observed market shares. Honka (2014) separates the nonprice utility of a firm into a portion due to advertising and an idiosyncratic portion, both of which are subsumed into α_j here. Three small insurers are excluded due to their limited geographic presence at the time of the data collection: American Family, Erie, and Mercury.

Price coefficent γ	0.3978
Inertia coefficient β	1.3375
Scale parameter μ of the Gumbel price distribution	1.4517
Search cost c	0.1663

Table 2.8.3: Additional parameter estimates from Model 0 of Honka (2014). Prices are in hundreds of dollars for six months of coverage.

Discount rate	Mean cost share of price
0.00	0.6520
0.06	0.6487
0.10	0.6461
Infinite	0.5672

Table 2.8.4: Summary of cost estimation results under alternative assumptions of how much firms discount future profits. This paper proceeds with an assumption of 6 percent (0.06) annual discounting.

	$\overline{r}_j$	$\frac{\overline{r}_j}{\overline{p}_j}$	Total surplus (mean)	2006-2007 market share (pct.)	Discounted share (pct.)	2018 share	Predicted 2018 share (pct.)
21st Century	4.01	0.64	1.46	3.05	7.68	–	7.88
AIG	4.38	0.65	1.59	4.29	9.18	–	9.40
Allstate	4.67	0.66	1.34	15.35	7.04	9.22	6.65
Farmers	3.87	0.62	1.74	5.33	10.81	4.27	11.07
Geico	3.47	0.59	1.28	19.95	6.64	13.45	6.05
GMAC	5.06	0.69	0.88	1.73	3.92	–	4.00
The Hartford	3.62	0.60	1.86	4.60	12.71	–	13.12
Liberty Mutual	4.74	0.67	1.68	5.13	10.14	4.79	10.37
Metlife	4.64	0.67	0.94	2.48	4.24	–	4.30
Nationwide	4.40	0.66	0.96	3.63	4.47	2.74	4.50
Progressive	4.08	0.63	1.54	13.46	8.77	11.01	8.53
Safeco	4.15	0.65	0.73	1.82	3.38	–	3.43
State Farm	4.66	0.66	1.06	14.29	5.06	17.07	4.67
Travelers	5.00	0.69	1.20	4.89	5.96	1.91	6.00

Table 2.8.5: Estimated marginal costs $\overline{r}_j$ and actual and projected market shares. The discounted market share is a sum of market shares from the data collection in 2006 until 50 years later, discounted at a 6 percent annual rate and normalized to sum to 100 percent. 2018 market shares are from NerdWallet (https://www.nerdwallet.com/blog/insurance/car-insurance-basics/largest-auto-insurance-companies/).

	Reported for	Year	Premiums ($B)	Claims (pct.)	Underwriting ratio	Combined ratio
Progressive	All insurance	2006	14.1	66.5	20.1	86.6
		2005	13.8	68.0	20.1	88.1
		2004	13.2	65.0	20.2	85.2
State Farm	Auto	2007	31.7	78.7	20.6	99.3
		2006	31.9	73.6	21.2	94.8
Geico	"primarily" auto	2006	11.1	70.1	18.0	88.1
		2005	10.1	70.6	17.3	87.9
		2004	8.9	71.3	17.8	89.1
Allstate	All insurance	2007	29.1	64.9	24.9	89.8
		2006	29.3	58.5	25.1	83.6
		2005	29.1	78.3	24.1	102.4
		2004	28.1	68.7	24.3	93.0
		2003	27.0	70.6	24.0	94.6

Table 2.8.6: Expenses of major insurers on claims and underwriting, as a share of revenue from premiums. For some insurers, data includes insurance products other than auto insurance. Source: Annual reports of the companies.

Case	c	β	Consumer surplus	Consumer surplus (excl. search costs and inertia)	Num. searches	Mean price paid	Mean price offer
Est.	0.17	1.34	8.93	8.65	1.82	6.25	6.60
1	0.17	0.67	8.31	8.79	1.93	6.54	6.83
2	0.17	0.00	7.82	8.66	2.00	6.85	7.07
3	2.00	1.34	2.55	4.70	1.06	5.28	5.29
4	1.50	1.34	4.14	5.50	1.13	5.41	5.40
5	1.00	1.34	5.92	6.90	1.31	5.68	5.70
6	0.50	1.34	7.69	8.16	1.58	6.10	6.23
7	0.10	1.34	9.29	8.85	2.00	6.13	6.61
8	0.05	1.34	9.60	9.02	2.28	5.98	6.61
9	0.00	1.34	9.97	9.22	14.00	5.70	6.62
10	0.00	0.00	9.15	9.15	14.00	6.04	6.95

Table 2.8.7: Results of counterfactually varying the search cost c and inertia preference β. The baseline price p_j is weighted by observed market shares of firms. The average number of searches with the estimated model differs from the average number of searches (2.96) observed by Honka (2014), indicating some weakness in the fit of Model 0.

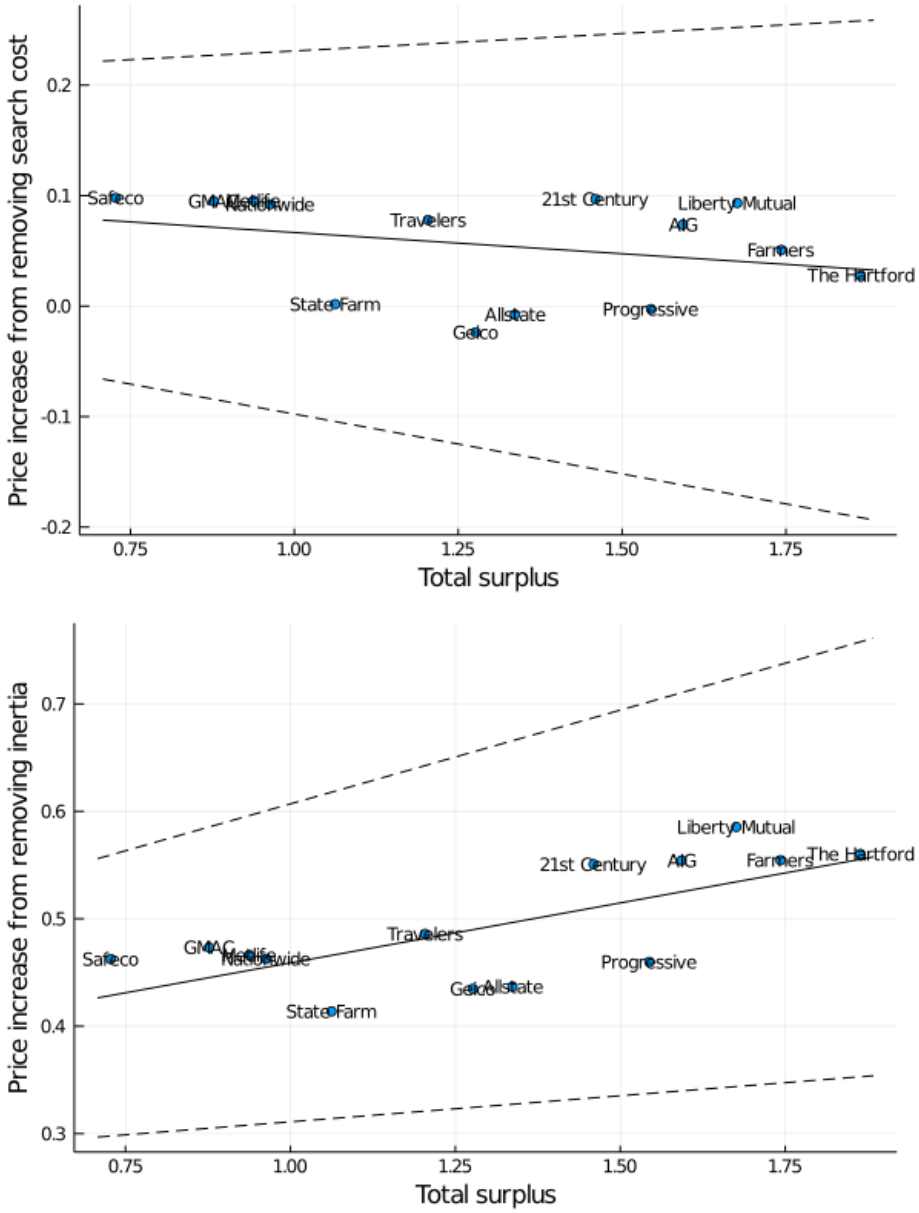

Figure 2.8.1: Firms' total surplus versus the increase in firm price choices when search or switching costs are removed. Linear trend with 95 percent confidence interval.

Appendix

2.A Price Derivatives of the Transition Matrix

The derivatives $\frac{dP_{kj'}}{d\overline{p}_j}$ are needed both to estimate marginal costs and to construct the Taylor approximation to the profit function for approximate best response iteration. To begin, Equation 2.8 is differentiated using the product rule.

$$\begin{aligned}
&\frac{dP_{kj'}}{d\overline{p}_j} \qquad (2.18)\\
&= \sum_K [\frac{d}{d\overline{p}_j} \Pr(K \text{ searched} | \varepsilon_{ijt}\ \forall j \neq j', W_k = 0)\\
&\times \Pr(j' \text{ accepted} | K \text{ searched}, \varepsilon_{ijt}\ \forall j \neq j', W_k = 0)]\\
&+ [\Pr(K \text{ searched} | \varepsilon_{ijt}\ \forall j \neq j', W_k = 0)\\
&\times \frac{d}{d\overline{p}_j} \Pr(j' \text{ accepted}, \varepsilon_{ijt}\ \forall j \neq j', W_k = 0)
\end{aligned}$$

The right hand side contains the search probability $Pr(K_i \text{ searched} | \varepsilon_{ijt}\ \forall j \neq j', W_k = 0)$, the acceptance probability $\Pr(j' \text{ accepted} | K \text{ searched}, \varepsilon_{ijt}\ \forall j \neq j', W_k = 0)$, and their derivatives with respect to the mean price offer chosen by firm j. The search probability is found by identifying the optimal search set as a function of $\varepsilon_{ij't}$ while simulating ε_{ijt} for all $j \neq j'$. The acceptance probability further simulates over price shocks $p_{ijt} - \overline{p}_j$ for all firms. These calculations are discussed in Section 2.5.

The derivative of the search probability $Pr(K_i \text{ searched} | \varepsilon_{ijt}\ \forall j \neq j', W_k = 0)$ is a function of how changes in prices affect the thresholds of $\varepsilon_{ij't}$ that separate the optimal search sets. Consider the threshold between search sets K and K'. This threshold is the value of $\varepsilon_{ij't}$ such that the consumer has the same expected utility from each search set:

$$EU(K | \varepsilon_{ij't}, \overline{p}) - EU(K' | \varepsilon_{ij't}, \overline{p}) = 0 \qquad (2.19)$$

Totally differentiate to obtain

$$d\varepsilon_{ij't}\frac{d}{d\varepsilon_{ij't}}(EU(K|\varepsilon_{ij't},\overline{p}) - EU(K'|\varepsilon_{ij't},\overline{p})) \\ +d\overline{p}_j\frac{d}{d\overline{p}_j}EU(K|\varepsilon_{ij't},\overline{p}) - EU(K'|\varepsilon_{ij't},\overline{p})) \tag{2.20}$$

$$\frac{d\varepsilon_{ij't}}{d\overline{p}_j} = \frac{\frac{d}{d\overline{p}_j}EU(K|\varepsilon_{ij't},\overline{p}) - EU(K'|\varepsilon_{ij't},\overline{p}))}{\frac{d}{d\varepsilon_{ij't}}(EU(K|\varepsilon_{ij't},\overline{p}) - EU(K'|\varepsilon_{ij't},\overline{p}))} \tag{2.21}$$

Equation 2.21 can be easily evaluated by observing that conditional on the search set K, an increase in $\overline{p}_{j'}$ affects only the utility of firm j'. The overall effect on expected utility equals the change in the utility of firm j' times the probability of firm j' being chosen conditional on search set K.

The derivative of the acceptance probability $\Pr(j' \text{ accepted}|K \text{ searched}, \varepsilon_{ijt}\,\forall j \neq j', W_k = 0)$ is more difficult to evaluate, and I turn to an automatic differentiation routine for ease and reliability (Julia's `ForwardDiff`). Differentiation begins with the same approach used to evaluate search probabilities. I simulate over ε_{ijt} for $j \neq j'$ and over $p_{ijt} - \overline{p}_j$ for all j. Firm j' is accepted if

$$\alpha_{j'} - \gamma\overline{p}_{j'} + \varepsilon_{ij't} > \alpha_j - \gamma\overline{p}_j + \varepsilon_{ijt}\,\forall j \neq j' \tag{2.22}$$

which I write as a cutoff on the remaining unsimulated shock $\varepsilon_{ij't}$ as

$$\varepsilon_{ij't} > \alpha_j - \gamma\overline{p}_j + \varepsilon_{ijt} - \alpha_{j'} + \gamma\overline{p}_{j'}\,\forall j \neq j' \tag{2.23}$$

The probability of acceptance (conditional on search) is

$$\Pr(\varepsilon_{ij't} > \alpha_j - \gamma\overline{p}_j + \varepsilon_{ijt} - \alpha_{j'} + \gamma\overline{p}_{j'}\,\forall j \neq j'|K \text{ searched}) \tag{2.24}$$

where the conditioning reflects that $\varepsilon_{ij't}$ is necessarily within a certain range given that the consumer chose to search the set K. This range too depends on firms' mean price offers p. Differentiation of Equation 2.24 is complex but is readily handled by an automatic differentiation program such as Julia's `ForwardDiff`.

Chapter 3

Generalized Linear Models for Demand Estimation with Lightly Aggregated Data

Researchers seeking to estimate structural demand models sometimes have the benefit of lightly aggregated data. Instead of total unit sales of a few broadly defined products, lightly aggregated data reports the allocation of unit sales across other levels such as within-product varieties, retailers, time periods, or (groups of) consumers. Options that represent the same product, time period, or other category may share some common effect in their utility. With many distinct options for purchase, an individual option's observed sales may be low and thus subject to sampling error.

As applied to lightly aggregated data, aggregate data demand estimation following BLP has some undesirable properties. First, BLP assume that the observed market shares are equal to the underlying choice probabilities. Apart from leading to false precision in statistical inference, this assumption in disaggregate data may cause the model parameters to overfit and capture chance variations in the choices of a few consumers. Further, there is no easy or accepted method of restricting the model parameters to avoid overfitting. Each option is allowed its own idiosyncratic fixed effect in the utility function, no matter how few observed sales are available for estimating this effect.

This paper proposes an alternative estimation strategy specifically for lightly aggregated data. Motivating this strategy is the observation that unit sales of each option follow a binomial distribution to be modeled by binomial regression. Binomial regression falls within the class of generalized linear models (GLM), which can be estimated in modern standard statistical software with far less programming and computation time than methods based on BLP. Further, binomial regression allows for restrictive fixed effect structures that

can lessen the risk of overfitting. The key disadvantage of binomial regression is that it is limited to the simple multinomial logit model rather than the mixed logit model of BLP. However, I show that the Salanié and Wolak (2019) approximation to mixed logit can be applied to binomial regression without substantially increasing the computational cost.

Section 3.2 reviews other approaches to disaggregate data used in the literature. Section 3.3 presents this paper's methods: (1) a new representation of the data generating process in which each observation of unit sales is a draw from a binomial distribution, (2) a restrictive multilevel structure of utility based on the researcher's understanding of which options share common factors in utility, and (3) an application of the Salanié and Wolak (2019) linearization of mixed logit. Section 3.4 presents three Monte Carlo simulations. The first simulation considers a simple binary in versus out choice and compares binomial regression to alternative methods for dealing with sampling error such as Gandhi, Lu and Shi (2017). In small markets with few sales, only binomial regression is successful at recovering the true parameter values. However, binomial regression requires that the outside option is always a "safe" product in the sense of Gandhi, Lu and Shi (2017), meaning that it is chosen by a large enough number of consumers that the sampling error in this number is negligible. The second and third simulations consider a richer choice environment with multiple products and multiple firms, with the third simulation demonstrating the Salanié and Wolak (2019) linearization of mixed logit. Compared to competing methods, this linearization accepts some bias in return for computational speed and ease of implementation.

3.2 Literature

One common approach to demand estimation with lightly aggregated data has been to apply the same BLP-style methods used for aggregate data. Each option is treated as having its own distinct mean utility, imposing no commonality among different options that belong to one or more of the same categories. Miravete et al. (2020) for instance model the Pennsylvania liquor market with product-market-time fixed effects. With their reported 312 products, 454 markets, and 22 months, the authors would estimate over 3 million fixed effects. Given that only 41 million bottles were sold in their data, each fixed effect may be estimated from about 14 sales on average—or fewer for less popular products, creating a high risk of overfitting the fixed effects to sampling error. Other papers that allow for entirely distinct mean utility values within the same category include Murry (2017) (for vehicle models, dealers, and model years) and Nevo (2000) (for cereal brands, cities, and time at a quarterly frequency). This approach is well founded when the data is sufficiently aggregated that sampling error is small and market shares are approximately equal to choice

probabilities. In this case, there is no need for binomial regression or for a restrictive multilevel fixed effect structure. However, the Salanié and Wolak (2019) linearization may still be useful in reducing computation time.

In some cases, researchers may have wished to estimate a BLP-style mixed logit model but were prevented from doing so by the disaggregation of their data. In a paper using product-store-week sales data, Mummalaneni et al. (2019) explain that although they control for various forms of observable heterogeneity in a multinomial logit model,

> [c]ontrolling for unobservable heterogeneity as in Berry et al. (1995) is difficult in this context due to the granularity of the data: computational feasibility is only plausible if we aggregate across products, across stores,or across weeks—this in turn would result in the loss of important variation in the data.

Simplifying to multinomial logit resolves the computational difficulties at some loss of realistic substitution patterns. However, it does not address the concern that some product-store-week options might have too few sales to reliably estimate a fully flexible fixed effect.

A second strand of work has supplemented disaggregate or lightly aggregated data with aggregate data in which sampling error is assumed to be negligible. Chintagunta and Dubé (2005) combine panel data on household grocery purchases with a larger aggregate dataset from stores. Demand is estimated by maximum likelihood on the household data, subject to the constraint that the model correctly predicts sales in the aggregate data. Berry et al. (2004) use disaggregate consumer-level survey data to add additional moments to generalized method of moments (GMM) estimation on aggregate data. Applications with multiple levels of differentiation (e.g. product, retailer, market) may not have a single level of aggregation to which these methods would apply.

A third strand, to which this paper belongs, probabilistically accounts for the sample noise inherent in small or disaggregate data. Gandhi, Lu and Shi (2017) observe that the quantity sold of a good j follows a binomial distribution: $q_j \sim \text{Binomial}(M, s_j)$, where M is the number of consumers in the market and s_j is the choice probability or asymptotic market share. The authors propose probabilistic bounds on s_j based on the realized q_j. A necessary assumption for Gandhi et al. (2017) is that some options are "safe", in the sense of having enough sales to accurately estimate their choice probabilities. Compared to the Gandhi et al. (2017) method, binomial regression has the advantage of better performance for low-sales products in small markets, and of easily accommodating multilevel fixed effects for products that share a common element of their utility due to membership in the same category. In the alcohol example of Miravete, Seim and Thurk (2020), one choice of category might be all red wines sold in a given time period, or at a given store. With appro-

priately defined categories, the researcher may be able to conclude more about the quality of each option than could be learned from the Gandhi et al. (2017) bounds constructed solely based on the option's own sales. The disadvantages of binomial regression compared to Gandhi et al. (2017) are (1) the need for Assumption 1 below, which is satisifed when the outside option is frequently chosen in every market and (2) substantial bias when extending to mixed logit via the Salanié and Wolak (2019) linearization.

3.3 Model and estimation

Consumer i purchases product j from retailer or dealer d in market t. Alternatively the consumer may purchase an outside option with mean utility normalized to zero. An inside option's utility is given by

$$u_{ijdt} = \delta_{jdt} + \varepsilon_{ijdt} \tag{3.1}$$

$$\text{where } \delta_{jdt} = \boldsymbol{H}_{jdt}\gamma + \xi_j^J + \xi_d^D \tag{3.2}$$

Here $\boldsymbol{H}_{jdt}$ is a vector of attributes of the particular combination j, d, t and potentially including attributes such as market demographics or travel time to the retailer. The multilevel fixed effect structure $\xi_j^J + \xi_d^D$ is less flexible compared to the more commonly estimated ξ_{jdt} but is therefore less susceptible to overfitting—provided, of course, that the multilevel structure approximates the true utility model. Other fixed effect specifications could also be accommodated in the binomial regression framework, e.g. adding a market fixed effect ξ_t^T.

The idiosyncratic error ε_{ijdt} is distributed Gumbel$(0, 1)$. The resulting choice probabilities are given by the standard multinomial logit formula

$$s_{jdt} = \frac{\exp(\delta_{jdt})}{1 + \sum_{j',d'} \exp(\delta_{j'd't})} \tag{3.3}$$

The product fixed effects ξ_j^J will generally be a function of observable product characteristics $\boldsymbol{X}_j$.

$$\xi_j^J = \boldsymbol{X}_j\beta + \zeta_j^J \tag{3.4}$$

I assume that the researcher observes M_t, q_{jdt}, $\boldsymbol{X}_j$, and instruments $\boldsymbol{Z}_j$ to account for any correlation between ζ_j^J and elements of $\boldsymbol{X}_j$.

3.3.1 Multinomial logit

Fix a particular option (j, d, t) and consider a consumer in market t who is known to be choosing either (j, d, t) or the outside option (which has choice probability s_{0t}). The conditional market share of (j, d, t) can be defined as

$$\tilde{s}_{jdt} := \frac{s_{jdt}}{s_{0t} + s_{jdt}} \tag{3.5}$$

$$= \frac{\frac{\exp(\delta_{jdt})}{1+\sum_{j',d'} \exp(\delta_{j'd't})}}{\frac{\exp(0)}{1+\sum_{j',d'} \exp(\delta_{j'd't})} + \frac{\exp(\delta_{jdt})}{1+\sum_{j',d'} \exp(\delta_{j'd't})}} \tag{3.6}$$

$$= \frac{\exp(\delta_{jdt})}{1 + \exp(\delta_{jdt})} \tag{3.7}$$

Define analogously the conditional market size $\tilde{M} = M_t(s_{0t} + s_{jdt})$. I approximate $\tilde{M}_t$ by the sum of unit sales $q_{0t} + q_{jdt}$, which is equal to $\tilde{M}_t$ in expectation.

Assumption 1. *There is negligible error in the approximation* $q_{0t} + q_{jdt} = M_t(s_{0t} + s_{jdt})$

Assumption 1 is weaker than the usual BLP assumption of $q_{jdt} = M_t s_{jdt}$. Rather than assuming that each option has enough observed sales to compute its choice probability, I assume that the outside option and each inside option together have enough sales to compute the sum of their choice probabilities. This assumption is most closely satisfied when both the unconditional market size M_t and the outside option share s_{0t} are large.[1]

Subjection to Assumption 1, (j, d, t) has unit sales distributed

$$q_{jdt} \sim \text{Binomial}(q_{0t} + q_{jdt}, \frac{\exp(\delta_{jdt})}{1 + \exp(\delta_{jdt})}) \tag{3.8}$$

Equation 3.8 relates the unobserved utility δ_{jdt} to observed sales through a standard distributional form: the binomial distribution with binary logit link function. Further, Equation 3.4 states that δ_{jdt} is a linear function of its parameters. These parameters can thus be estimated by the standard technique of binomial regression.

I estimate Equation 3.8 in R using the package `glmmboot`, a fast bootstrap maximum likelihood algorithm for generalized linear mixed model (GLMM) regression that is capable of profiling out one level of fixed effects. For the fastest solution, the level with the most categories should be selected to be profiled out while other levels of fixed effects are included as factor variables. Similar functionality is available in other software (e.g. the

[1] In applications where there is no outside option or where the outside option is rarely chosen, a commonly chosen product can play the role of the outside option by having its utility normalized to zero.

`melogit` command in Stata, or the `MixedModels` package in Julia). Following estimation of Equation 3.8, Equation 3.4 can be estimated by two-stage least squares (2SLS) to recover an estimate of $\boldsymbol{\beta}$.

The estimation copes naturally with the possibility of a zero empirical market share for a single option $q_{jdt} = 0$. An empirical market share of zero prevents estimation of the standard aggregate data logit or BLP model, unless a correction is applied as in Gandhi et al. (2017). Yet a similar problem occurs in binomial regression, albeit at a higher level of aggregation. If any fixed effect ξ_d^D or ξ_j^J has $q_{jdt} = 0$ for every option in which it occurs, then the maximum likelihood estimate of that fixed effect is $-\infty$. Apart from being unreasonable, such a result shuts down any analysis of how product attributes relate to ξ_j^J. I therefore assume that

Assumption 2.
For all j, $\exists(d,t)$ *such that* $q_{jdt} > 0$
For all d, $\exists(j,t)$ *such that* $q_{jdt} > 0$

Ideally one would hope for a much stronger condition: that each fixed effect is estimated from so many observed purchases that there is little sampling error. To the extent that sampling error occurs, it is accounted for in estimation and decreases the estimated precision of the results. The researcher can respond to excessive sampling error by collecting additional data or imposing more commonality in fixed effects by combining categories.

3.3.2 Extension to mixed logit

The mixed logit model allows consumers to differ in their valuations of product characteristics, leading to more flexible and realistic substitution patterns. For a mixed logit model, modify Equation 3.4 by inserting a random vector $\boldsymbol{v}_i \sim \text{Normal}(0, \boldsymbol{\Sigma})$, independent across consumers.

$$u_{ijdt} = \delta_{jdt} + \boldsymbol{X}_j \boldsymbol{v}_i + \varepsilon_{ijdt} \tag{3.9}$$

The resulting market shares are

$$s_{jdt} = \int_{\boldsymbol{v}_i} \frac{\exp(\delta_{jdt} + \boldsymbol{X}_j \boldsymbol{v}_i)}{1 + \sum_{j',d'} \exp(\delta_{j'd't} + \boldsymbol{X}_{j'} \boldsymbol{v}_i)} \tag{3.10}$$

The traditional approach is to evaluate this integral by simulation. Market shares depend nonlinearly on $\boldsymbol{\Sigma}$, requiring a nonlinear optimization to estimate $\boldsymbol{\Sigma}$ by generalized method of moments (GMM). Moreover, the researcher must solve at every optimization

iteration for the fixed effect values that best explain the observed market shares given $\boldsymbol{\Sigma}$. BLP solve this subproblem using a nested fixed point algorithm to fit a fully flexible δ_{jdt} for each option. This approach is slow and subject to numerical error, leading Dubé, Fox and Su (2012) and Lee and Seo (2015) to propose improvements: the former reframing the contraction mapping as a constrained optimization, the latter linearizing the dependence of market shares on δ_{jdt}. All of these approaches are significantly more complex and computation-intensive than binomial regression. Moreover, they do not account for sampling error (without further modification), and cannot readily accommodate multilevel fixed effects to prevent overfitting.

Salanié and Wolak (2019) observe that the transformation $\log(\frac{s_{jdt}}{s_{0t}})$ of the mixed logit market shares 3.10 can be approximated by a Taylor series in $\boldsymbol{\Sigma}$. Their paper develops a 2SLS estimation strategy based on this linearization, circumventing the computational complexity of BLP-style methods at the cost of some approximation error. The remainder of this subsection shows how the same linearization allows mixed logit models to be estimated by binomial regression.

Let m index the various product characteristics in $\boldsymbol{X}_j$. As implied by Theorem 2 in Salanié and Wolak (2019), the Taylor series of $\log(\frac{s_{jdt}}{s_{0t}})$ is

$$\log(\frac{s_{jdt}}{s_{0t}}) = \zeta_j^J + \xi_d^D + \boldsymbol{X}_j\boldsymbol{\beta} + \sum_m \Sigma_{mm} K_{mm}^{jdt} + \sum_{m<n} \Sigma_{mn}(K_{mn}^{jdt} + K_{nm}^{jdt}) + O(||\Sigma||^{k/2}) \tag{3.11}$$

where $\boldsymbol{K}$ is a set of artificial regressors to be defined as functions of product characteristics and market shares. First, for each market t define the market share weighted covariate vector $\boldsymbol{e}_t = \sum_j s_{jdt}\boldsymbol{X}_{jdt}$. Let $\mathcal{I}$ be the set of pairs (m, n) of characteristics where Σ_{mn} is not assumed to be zero. For each option (j, d) in t and for each pair of indices $(m, n) \in \mathcal{I}$, complete the definition of $\boldsymbol{K}$ as

$$K_{mn}^{jdt} = (\frac{X_{jdtm}}{2} - e_{tm})X_{jdtn} \tag{3.12}$$

To apply the Salanié and Wolak (2019) linearization to binomial regression, begin by defining

$$\overline{\delta}_{jdt} = \zeta_j^J + \xi_d^D + \sum_m \Sigma_{mm} K_{mm}^{jdt} + \sum_{m<n} \Sigma_{mn}(K_{mn}^{jdt} + K_{nm}^{jdt}) + O(||\Sigma||^{k/2}) \tag{3.13}$$

Equation 3.11 then implies

$$\frac{s_{jdt}}{s_{0t}} = \exp(\overline{\delta}_{jdt}) \tag{3.14}$$

$$\frac{s_{0t}}{s_{jdt}} = \exp(-\overline{\delta}_{jdt}) \tag{3.15}$$

$$\frac{s_{0t} + s_{jdt}}{s_{jdt}} = 1 + \exp(-\overline{\delta}_{jdt}) \tag{3.16}$$

$$\frac{s_{jdt}}{s_{0t} + s_{jdt}} = \frac{1}{1 + \exp(-\overline{\delta}_{jdt})} \tag{3.17}$$

$$\tilde{s}_{jdt} = \frac{\exp(\overline{\delta}_{jdt})}{1 + \exp(\overline{\delta}_{jdt})} \tag{3.18}$$

which gives the probability of choosing (j, d, t) conditional on choosing either (j, d, t) or the outside option. As in Equation 3.8 for the multinomial logit, unit sales have a binomial distribution

$$q_{jdt} \sim \text{Binomial}(q_{0t} + q_{jdt}, \frac{\exp(\overline{\delta}_{jdt})}{1 + \exp(\overline{\delta}_{jdt})}) \tag{3.19}$$

where $q_{0t} + q_{jdt}$ approximates $M_t(s_{0t} + s_{jdt})$ by Assumption 1. It is thus possible to estimate the parameters $\boldsymbol{\beta}, \boldsymbol{\Sigma}, \boldsymbol{\zeta}^J, \boldsymbol{\xi}^D$ by binomial regression, where Equation 3.13 now forms the input to the logit link function. This procedure is applied in the last Monte Carlo simulation of Section 3.4. Importantly, the artificial regressors $\boldsymbol{K}$ are endogenous due to their construction from market shares and will typically require instruments. As proposed by Salanié and Wolak (2019), the instruments for $\boldsymbol{K}$ and those for any endogenous elements of $\boldsymbol{X}$ may be functions of the same set of variables $\boldsymbol{Z}$, and of exogenous elements of X. In this paper, I assume for ease of exposition that $\boldsymbol{X}$ is exogenous and use only the portion of the Salanié and Wolak (2019) instruments based solely on $\boldsymbol{X}$.

Although binomial regression accounts for sampling error in q_{jdt} as an outcome, it does not account for the sampling error inherent in using the noisy empirical market shares s_{jdt} to construct the regressors $\boldsymbol{K}$. However, the construction of $\boldsymbol{K}$ sums over the market shares of all products in the market, potentially reducing the impact of sampling error.[2]

[2]If the sampling error in $\boldsymbol{K}$ is expected to be severe, it may be advisable to replace the market shares with Bayesian estimates of their asymptotic shares following Gandhi et al. (2017).

3.4 Monte Carlo simulations

Three Monte Carlo simulations demonstrate the value of binomial regression for lightly aggregated data: to account for sampling error (including the case of zero observed sales), to estimate multilevel fixed effects, and to combine these benefits with the computationally efficient Salanié and Wolak (2019) approach to mixed logit estimation.

The first simulation considers a simple binary choice between a single inside good with utility $\beta + \varepsilon_{i1t}$ and an outside option with utility ε_{i0t}. The inside good is of poor quality with $\beta = -9$, implying that its unit sales will be subject to substantial sampling error when the market size M_t is small. Depending on the simulation, the market size ranges from 100 to 1,000,000.

Using this simulated data, I compare binomial regression with two competing corrections for sampling error, the first of which is flawed but often used in empirical work for its simplicity. Rather than fully account for the sampling error in market shares, the researcher only addresses the zero market shares that would otherwise block estimation. OLS logit estimation defines the dependent variable $\log(\frac{s_{jdt}}{s_{0t}})$, which is undefined if s_{jdt} is taken to be its observed value of zero. Reallocating one sale from the outside option allows the estimation to proceed via the usual OLS regression

$$\log(\frac{s_{jdt}}{s_{0t}}) = \boldsymbol{X}_j\beta + \xi_{jdt} \tag{3.20}$$

Gandhi, Lu and Shi (2017) observe that this naive correction introduces bias, as do similar methods such as dropping options with zero observed sales. They propose instead a set of probabilistic bounds on the true choice probabilities, with an estimation strategy that seeks to minimize departures from these bounds. The bounds on each option's mean utility δ_{jt} are defined as

$$\delta^u_{jt} = \log(\frac{q_{jt} + l_u}{M_t}) - \log(\tilde{s}_{0t}) \tag{3.21}$$

$$\delta^\ell_{jt} = \log(\frac{q_{jt} + l_\ell}{M_t}) - \log(\tilde{s}_{0t}) \tag{3.22}$$

where $\tilde{s}_{0t}$ is the "Laplace share" of the outside good, shifted toward the interior of the $[0, 1]$ interval.[3] Defining J_t as the number of products in market t (excluding the outside

[3] Specifically, $\tilde{s}_{0t} = \frac{q_t+1}{q_t+J_t+M_t}$.

Market size		Mean β_0	s.e.
100	Binomial	-9.07	0.0324
	OLS	-2.20	0.0000
	Gandhi	[-709.76, -3.12]	[0.0275, 0.0001]
1,000	Binomial	-9.02	0.0086
	OLS	-4.60	0.0000
	Gandhi	[-704.41, -5.29]	[0.0738, 0.0001]
10,000	Binomial	-9.00	0.0027
	OLS	-6.90	0.0001
	Gandhi	[-633.41, -7.46]	[0.2062, 0.0003]
100,000	Binomial	-9.00	0.0008
	OLS	-8.90	0.0004
	Gandhi	[-215.15, -8.88]	[0.3152, 0.0006]
1,000,000	Binomial	-9.00	0.0003
	OLS	-9.04	0.0003
	Gandhi	[-9.04, -9.00]	[0.0003, 0.0003]

Table 3.4.1: Comparison of three estimation methods for a binary choice problem: binomial regression, OLS (reallocating one sale from the outside to the inside option if the inside option has zero sales), and the Gandhi et al. (2017) bounds method. The true β_0 is -9. Data: 100 simulations of 10,000 markets.

option),

$$\tilde{s}_t = \frac{q_{0t} + 1}{M_t + J_t + 1} \tag{3.23}$$

The scalars l_u and l_ℓ are chosen as functions of M_t and are constructed to guarantee proper coverage of the bounds whatever the true choice probability s_{jt}. Details of this construction are given in Gandhi et al. (2017). The result is that l_ℓ is a small positive number while l_u ranges from 0.405 for $M_t = 10$ to 0.509 for $M_t = 10,000$ (conditional on a binary choice environment). The Gandhi et al. (2017) estimator minimizes a function of the deviations from these bounds.

Table 3.4.1 presents simulation results from the extreme case of $\beta_0 = -9$, implying that about 1 in 1,000 consumers purchases the product. Binomial regression outperforms the naive OLS solution, with the disparity decreasing as market size increases. Comparisons to the Gandhi method are less straightforward as it identifies sets rather than points. While the set estimates did in every case contain the true parameter value, they were too broad to be of practical use in even fairly large markets ($M_t = 100,000$).

The disparity in Table 3.4.1 can be explained by contrasting the Gandhi method's use of moment inequalities with the maximum likelihood approach of binomial regression. Fix-

ing $M = 10$, the Gandhi approach interprets a market with zero sales as imposing an upper bound on β_0 of $\delta^u = -3.12$. Observing many such markets in the Gandhi method provides further evidence that $\beta_0 \leq -3.12$, but does not serve to tighten this bound. Binomial regression instead takes full advantage of these repeated observations. It would be unlikely to observe zero sales in almost all markets if β_0 were as high as -3.12, and a higher likelihood can be obtained from the true $\beta_0 = -9$. Binomial regression thus combines information from across markets in the form of likelihood contributions, allowing for estimation on small markets that provide little useful information individually.

Gandhi et al. (2017) present Monte Carlo simulations in which their method produces correct estimates despite most options having zero sales. They assume a data generating process in which 99 percent of options are of uniform and low quality while the remaining 1 percent are of variable and high quality. It is the presence of the high-quality products, and the variation in their quality, that makes the estimation successful. Because the markets are large ($M_t = 10,000$), the high-quality products are "safe" in Gandhi et al.'s terminology: their actual market shares are roughly equal to their expected market shares. As Table 3.4.1 shows, disaggregate data with small markets may not have any "safe" products, motivating the use of binomial regression to fully leverage the information from "risky" products across many markets.

The second and third simulations turn to the case of a multiproduct market, in which a few products appear repeatedly as options across many markets. In the second simulation, a set of 200 national brand products are allocated randomly across 10 retailers, with each retailer offering 20 of these products. The same national brand product may be offered by multiple retailers. Each dealer contributes utility $\xi_d^D \sim \text{Uniform}(-1, 1)$ to all of its options. To add a difficulty to the estimation that motivates the use of a dealer fixed effects, dealers with $\xi_d^D > 0$ stock only those national products with $X_1 > 0$. In addition, each retailer offers (with equal probability) either 0, 1, or 2 generic products specific to the retailer. Each of 1,000 markets (of 100 consumers each) has access to two random retailers. Utility follows the exposition in Section 3.3 and is the sum of utility ξ_d^D from the retailer, utility $\boldsymbol{X}_j\beta$ from researcher-observed product characteristics, utility $\zeta_j^J \sim \text{Uniform}(-1, 1)$ from an unobserved product characteristic, and an idiosyncratic shock $\varepsilon_{ijdt} \sim \text{Gumbel}(0, 1)$. The product characteristics are a constant (associated with $\beta_0 = -5$) and a scalar product characteristic (with $\beta_1 = 1$). The third simulation additionally assumes a $\text{Normal}(0, 1)$ random coefficient on the characteristic. Notably, the third simulation also replaces the binary X_1 with one distributed $\text{Normal}(0, 1)$. This revision to X_1 is necessary due to the Salanié and Wolak (2019) requirement of instruments for $\boldsymbol{K}$. The subset $\{X_1, X_1^2, X_1^3\}$ of the Salanié and Wolak (2019) instruments would be repetitive with a binary X_1, implying

	$\hat{\beta}_0$	$\hat{\beta}_1$
Estimate (Model FE)	-5.02	1.01
s.e.	(0.021)	(0.0115)
Estimate (Characteristics)	-4.85	1.00
s.e.	(0.0202)	(0.0145)
Estimate (Gandhi, fixing β_0)	-5.00	[-241, 1.73]
Estimate (Gandhi, fixing β_1)	[-247, -4.27]	1.00

Table 3.4.2: Results of three estimation methods for a multinomial logit choice process. The true coefficients are $\beta_0 = -5$, $\beta_1 = 1$. The first specification ("Model FE") estimates a binomial regression of sales on product and retailer fixed effects, then regresses the product fixed effects on product characteristics to obtain $\hat{\beta}_0$, $\hat{\beta}_1$. The second specification ("Characteristics") estimates a binomial regression of sales on product characteristics and retailer fixed effects, effectively ignoring the impact of idiosyncratic product utility ζ_j^J. The Gandhi estimates show a cross through the set identified by the method of Gandhi et al. (2017), with intersection at the true parameters. Data: 1,000 markets of 100 consumers each. Full details of the data generating process are in the text.

that the model would not be identified without additional excluded instruments as Salanié and Wolak (2019) assume are available.

Table 3.4.2 shows the results of the second simulation, which reliably recovers the coefficients given a sufficiently large sample. The main specification proceeds by a binomial regression of sales on product and retailer fixed effects, followed by a linear regression of the product fixed effects on product characteristics to obtain the parameters of interest. Also shown in Table 3.4.2 is a faster but less accurate alternative: a single binomial regression of sales on product characteristics and retailer fixed effects.

Table 3.4.2 also shows the results of the Gandhi method, which now produces two-dimensional set estimates $(\hat{\beta}_0, \hat{\beta}_1)$. In every simulation, these sets included the true parameters. To summarize the Gandhi set estimates, I trace a cross over the identified set with intersection at the true parameters. That is, I find (1) $\hat{\beta}_0$ such that $(\hat{\beta}_0, \beta_1)$ is within the set estimate and (2) $\hat{\beta}_1$ such that $(\beta_0, \hat{\beta}_1)$ is within the set estimate. With a small market size, I find that this cross is large (and so the identified set is larger still). As in the binary logit case of Table 3.4.1, the Gandhi approach relies on the presence of large markets in which at least some options are "safe", i.e. likely to have market shares approximately equal to their expected market shares.

The third simulation (Table 3.4.3) shows that binomial regression can be adapted to study the mixed logit model using the approximation of Salanié and Wolak (2019). As suggested by Salanié and Wolak (2019), the estimates $\hat{\sigma}^2$ are biased toward zero, i.e. toward the multinomial logit model. Nonetheless, this bias may be acceptable if computational

σ^2	Market size	$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\sigma}^2$
0.4	100	-4.865	1.013	0.286
		(0.020)	(0.009)	(0.018)
0.2	100	-4.881	1.011	0.178
		(0.019)	(0.010)	(0.018)
0.4	100,000	-4.827	1.004	0.270
		(0.018)	(0.009)	(0.017)
0.2	100,000	-4.888	1.004	0.162
		(0.020)	(0.008)	(0.016)
0.1	100,000	-5.006	1.000	0.090
		(0.017)	(0.000)	(0.000)
0.05	100,000	-4.987	1.000	0.047
		(0.017)	(0.000)	(0.000)

Table 3.4.3: Results of binomial regression for mixed logit, using the Salanié and Wolak (2019) linearization. The data generating process is as in Table 3.4.2 except that (1) there is a normally distributed random coefficient on X_1, with variance σ^2 and (2) the product characteristic X_1 is distributed Normal$(0, 1)$ instead of DiscreteUniform$(0, 1)$.

concerns or an absence of "safe" high-sales options prevents the researcher from implementing an exact mixed logit model. Some performance improvements may be possible without resorting to full mixed logit estimation. For example, Salanié and Wolak (2019) propose a bias correction based on the third and fourth-order terms of their linearization. The market shares used to construct $\boldsymbol{K}$ could also be corrected by the Bayesian methods discussed in Gandhi et al. (2017). Alternatively, the full Gandhi et al. (2017) method could be implemented to generate estimates that are likely to be less biased provided that some safe products are available to achieve identification.

3.5 Conclusion

Binomial regression is a standard approach to true individual-level microdata, in which case it becomes logit regression. This paper shows that binomial regression remains useful when consumer choices are aggregated, but not to the extent that a law of large numbers implies that market shares are equal to choice probabilities.

For multinomial logit demand, binomial regression is equivalent to maximum likelihood estimation as long as Assumption 1 holds. Binomial regression treats each option's sales as independent within a market (conditional on parameters), and therefore cannot incorporate mixed logit demand directly in a way that corresponds to maximum likelihood. However, the Salanié and Wolak (2019) linearization can approximate the realism of mixed

logit demand, without any major sacrifice of speed. In this way, binomial regression seeks a balance between speed and realism while addressing a third, increasingly frequent requirement of small-sample robustness.

Conclusion

This dissertation leverages disaggregate data to address issues of price discrimination and consumer search and switching, as well as methods for demand estimation. Price discrimination plays a substantial role in the mortgage market, suggesting a need for consumers to negotiate effectively rather than simply broaden their search to additional lenders. In the auto insurance market, search may in aggregate fail to benefit consumers as firms increase their prices to take advantage of the product differentiation revealed by search. Chapter 3 turns to methodology and shows that binomial regression may be used to estimate a logit demand model using lightly aggregated data where large-market assumptions fail to apply. If one is willing to accept some approximation error, binomial regression can also be used to estimate a mixed logit model without the usual computational difficulties.

BIBLIOGRAPHY

Agarwal, S., Grigsby, J., Hortacsu, A., Matvos, G., Seru, A., Yao, V., 2017. Search and Screening in Credit Markets. Working Paper .

Aguirregabiria, V., Ho, C.Y., 2012. A dynamic oligopoly game of the US airline industry: Estimation and policy experiments. Journal of Econometrics 168, 156 – 173. The Econometrics of Auctions and Games.

Alexandrov, A., Koulayev, S., 2017. No Shopping in the U.S. Mortgage Market: Direct and Strategic Effects of Providing Information. CFPB Working Paper .

Allen, J., Clark, R., Houde, J.F., 2014. The Effect of Mergers in Search Markets: Evidence from the Canadian Mortgage Industry. American Economic Review 104, 3365–96.

Allen, J., Clark, R., Houde, J.F., 2019. Search Frictions and Market Power in Negotiated-Price Markets. Journal of Political Economy Forthcoming. https://doi.org/10.1086/701684.

Anderson, S.P., Renault, R., 1999. Pricing, Product Diversity, and Search Costs: A Bertrand-Chamberlin-Diamond Model. The RAND Journal of Economics 30, 719–735.

Argento, R.B., Brown, L.M., Koulayev, S., Li, G., Myhre, M., Pafenberg, F., Patrabansh, S., 2019. First-Time Homebuyer Counseling and the Mortgage Selection Experience in the United States: Evidence from the National Survey of Mortgage Originations. CityScape 21.

Avery, R.B., Bilinski, M.F., Bucks, B.K., Chai, C., Chow, M., Clement, A., Critchfield, T., Frumkin, S., Keith, I.H., Mohamed, I.E., Pafenberg, F.W., Patrabansh, S., Schultz, J.D., Wood, C.E., 2017. A Profile of 2014 Mortgage Borrowers: Statistics from the National Survey of Mortgage Originations. NMDB Technical Report Series .

Avery, R.B., Borzekowski, R., 2019. Guest editors' introduction: National Survey of Mortgage Originations. Cityscape 21.

Baldick, R., Hogan, W.W., 2004. Polynomial Approximations and Supply Function Equilibrium Stability. Working paper .

Bartlett, R., Morse, A., Stanton, R., Wallace, N., 2018. Consumer-Lending Discrimination in the Era of FinTech. Working paper .

Berry, S., Haile, P.A., 2020. Nonparametric Identification of Differentiated Products Demand Using Micro Data. Working paper .

Berry, S., Levinsohn, J., Pakes, A., 1995. Automobile Prices in Market Equilibrium. Econometrica 63, 841–90.

Berry, S., Levinsohn, J., Pakes, A., 2004. Differentiated products demand systems from a combination of micro and macro data: The new car market. Journal of Political Economy 112, 68–105.

Bhutta, N., Fuster, A., Hizmo, A., 2018. Paying Too Much? Price Dispersion in the US Mortgage Market. Working Paper .

Bhutta, N., Hizmo, A., 2019. Do Minorities Pay More for Mortgages? Working Paper .

Bucks, B., Critchfield, T., Singer, S., 2019. National Survey of Mortgage Originations Survey Data on your Home Loan Toolkit. CityScape 21.

Cai, Q., Shahdad, S., 2015. What is the Mortgage Shopping Experience of Today's Homebuyer? Lessons from Recent Fannie Mae Acquisitions .

Chade, H., Smith, L., 2006. Simultaneous Search. Econometrica 74, 1293–1307. https://onlinelibrary.wiley.com/doi/pdf/10.1111/j.1468-0262.2006.00705.x.

Chintagunta, P., Dubé, J.P., 2005. Estimating an SKU-level Brand Choice Model Combining Household Panel Data and Store Data. Journal of Marketing Research XLII, 368–379.

Critchfield, T., Dey, J., Mota, N., Patrabansh, S., 2019. Mortgage Experiences of Rural Borrowers in the United States. CityScape 21.

Deltas, G., Li, Z., 2018. Free Riding on the Search of Others: Information Externalities in the Mortgage Industry. Working paper .

Diamond, P., 1971. A model of price adjustment. Journal of Economic Theory 3, 156 – 168.

Dubé, J.P., Fox, J.T., Su, C.L., 2012. Improving the numerical performance of static and dynamic aggregate discrete choice random coefficients demand estimation. Econometrica 80, 2231–2267.

Dubois, P., Perrone, H., 2018. Price Dispersion and Informational Frictions: Evidence from Supermarket Purchases .

Economides, N., 1989. Quality variations and maximal variety differentiation. Regional Science and Urban Economics 19, 21 – 29.

Ellison, G., Wolitzky, A., 2012. A search cost model of obfuscation. The RAND Journal of Economics 43, 417–441. https://onlinelibrary.wiley.com/doi/pdf/10.1111/j.1756-2171.2012.00180.x.

Facchinei, F., Fischer, A., Piccialli, V., 2009. Generalized Nash equilibrium problems and Newton methods. Math. Program. 117, 163–194.

Fang, L., Munneke, H.J., 2017. Gender equality in mortgage lending. Real Estate Economics .

Federal Reserve Board, 2009. Design and Testing of Truth-in-Lending Disclosures for Closed-End Mortgages .

Gandhi, A., Lu, Z., Shi, X., 2017. Estimating demand for differentiated products with zeroes in market share data. Working paper .

Govindan, S., Wilson, R., 2003. A global Newton method to compute Nash equilibria. Journal of Economic Theory 110, 65–86.

Haan, M.A., Moraga-González, J.L., Petrikaite, V., 2017. A Model of Directed Consumer Search. CEPR Discussion Paper .

Honka, E., 2014. Quantifying search and switching costs in the US auto insurance industry. The RAND Journal of Economics 45, 847–884. https://onlinelibrary.wiley.com/doi/pdf/10.1111/1756-2171.12073.

Honka, E., Chintagunta, P., 2017. Simultaneous or Sequential? Search Strategies in the U.S. Auto Insurance Industry. Marketing Science 36, 21–42. https://doi.org/10.1287/mksc.2016.0995.

Hortacsu, A., Syverson, C., 2004. Product Differentiation, Search Costs, and Competition in the Mutual Fund Industry: A Case Study of S&P 500 Index Funds. The Quarterly Journal of Economics 119, 403–456. http://oup.prod.sis.lan/qje/article-pdf/119/2/403/5351686/119-2-403.pdf.

Hotelling, H., 1929. Stability in competition. The Economic Journal 39, 41–57.

Klemperer, P., 1987. The competitiveness of markets with switching costs. The RAND Journal of Economics 18, 138–150.

Koulayev, S., 2014. Search for differentiated products: identification and estimation. The RAND Journal of Economics 45, 553–575. https://onlinelibrary.wiley.com/doi/pdf/10.1111/1756-2171.12062.

Lacko, J., Pappalardo, J., 1991. Improving consumer mortgage disclosures. Federal Trade Commission Bureau of Economics Staff Report .

Lee, J., Seo, K., 2015. A computationally fast estimator for random coefficients logit demand models using aggregate data. The RAND Journal of Economics 46, 86–102. https://onlinelibrary.wiley.com/doi/pdf/10.1111/1756-2171.12078.

Li, S., Başar, T., 1987. Distributed algorithms for the computation of noncooperative equilibria. Automatica 23, 523 – 533.

Matejka, F., McKay, A., 2015. Rational Inattention to Discrete Choices: A New Foundation for the Multinomial Logit Model. American Economic Review 105, 272–98.

McFadden, D., 1974. The Measurement of Urban Travel Demand. Journal of Public Economics 3, 303–328.

Miravete, E.J., Seim, K., Thurk, J., 2020. One Markup to Rule Them All: Taxation by Liquor Pricing Regulation. American Economic Journal: Microeconomics 12, 1–41.

Moraga-González, J.L., Sándor, Z., Wildenbeest, M.R., 2015. Consumer search and prices in the automobile market. Working Paper .

Moraga-González, J.L., Sándor, Z., Wildenbeest, M.R., 2017. Prices and heterogeneous search costs. The RAND Journal of Economics 48, 125–146. https://onlinelibrary.wiley.com/doi/pdf/10.1111/1756-2171.12170.

Mummalaneni, S., Wang, Y., Chintagunta, P.K., Dhar, S.K., 2019. Product placement effects on store sales: Evidence from consumer packaged goods. Working Paper .

Murry, C., 2017. Advertising in vertical relationships: An equilibrium model of the automobile industry. Working Paper .

Murry, C., Zhou, Y., 2016. Consumer search and automobile dealer co-location. Working Paper .

National Association of Insurance Commissioners: Casualty Actuarial and Statistical Task Force, 2015. Price Optimization White Paper .

Nevo, A., 2000. Mergers with differentiated products: The case of the ready-to-eat cereal industry. The RAND Journal of Economics 31, 395–421.

Pires, T., 2018. Measuring the effects of search costs on equilibrium prices and profits. International Journal of Industrial Organization 60, 179 – 205.

Redmer, C., 2019. Perceptions and expectations of mortgage borrowers: New evidence from the national survey of mortgage originations. CityScape 21.

Rubner, Y., Tomasi, C., Guibas, L.J., 1998. A metric for distributions with applications to image databases, in: Computer Vision, 1998. Sixth International Conference on, IEEE. pp. 59–66.

Salanié, B., Wolak, F.A., 2019. Fast, "robust", and approximately correct: Estimating mixed demand systems. NBER Working Paper .

Varner, M., Sankin, A., 2020. Why you may be paying too much for your car insurance. Consumer Reports .

Weitzman, M., 1979. Optimal search for the best alternative. Econometrica 47, 641–54.

von Weizsäcker, C.C., 1984. The costs of substitution. Econometrica 52, 1085–1116.

Wolinsky, A., 1986. True monopolistic competition as a result of imperfect information. The Quarterly Journal of Economics 101, 493–511.

Woodward, S.E., Hall, R.E., 2012. Diagnosing Consumer Confusion and Sub-Optimal Shopping Effort: Theory and Mortgage-Market Evidence. American Economic Review 102.