The Costs of Free Entry: Evidence from Real Estate Brokers in Greater Boston*

Panle Jia† Parag A. Pathak‡ Alan Genz§

This version: October 2010

Preliminary and Incomplete

Abstract

This paper studies the real estate brokerage industry in Greater Boston using a panel data set that covers about 260,000 properties and more than 10,000 agents from 1999 to 2007. We establish several stylized facts regarding entry and exit among real estate agents. First, there is a great deal of turnover: more than 12% of active real estate agents have entered the profession within the previous year, while about 10% of active agents exit each year. Second, patterns of entry and exit depend on market and time-series trends: agents are more likely to enter in markets and during time periods of housing price appreciation. Third, agents with three or more years of experience earn 60% more than agents with fewer than three years of experience and are 35% less likely to exit. These facts motivate a structural econometric model which we use to measure the welfare implications of free entry under the current fixed commission structure. The paper also develops a new method that allows estimation and counter-factual analysis for high-dimensional dynamic models. Our counter-factual results suggest that if commissions in later years had remained the same as in 1998, agent entry would be reduced by 51%, the exit rate would more than double, and social savings would amount to about 30% of industry revenues.

*We would like to thank Paul Asquith, Lanier Benkard, Steve Berry, Xiaohong Chen, JP Dube, Ariel Pakes, Yi Qian, and seminar participants at John Hopkins, and Yale for their helpful comments. Comments are welcome.

†Department of Economics, M.I.T. and NBER. Email: pjia@mit.edu. Corresponding author.

‡Department of Economics, M.I.T. and NBER. Email: ppathak@mit.edu.

§Department of Mathmatics, Washington State University. Email: genz@math.wsu.edu.

1 Introduction

Buying or selling a home is one of the most important financial decisions for a large majority of households in the United States. In 2007, 68% of households owned their homes, more than a quarter of national wealth was held in residential real estate, and there were 6.4 million sales of existing homes, according to estimates from the Department of Housing and Urban Development.[1,2] Surveys indicate that an overwhelming majority of homes are sold with the aide of a licensed real estate agent or broker.[3] The National Association of Realtors (NAR), the largest professional organization of real estate agents in the United States that represents more than half of all licensed agents, estimates that nationally almost 80% of residential real estate transactions involve a realtor. Brokers' commissions on the sale of real estate properties amounted to roughly $96 billion in 2004 and $102 billion in 2006.[4]

The real estate brokerage industry has two unique features. First, unlike many other industries, prices that realtors charge, their commission rates, are similar across different regions and have remained more or less unchanged for the past several decades.[5] This has been documented in a number of studies (see e.g., Hsieh and Moretti (2003) and Levitt and Syverson (2008)). Generally speaking, commissions range from 5% to 6% of a property's transaction price. A fixed commission implies that a property that sells because of strong housing market conditions and little agent effort generates the same commissions for an agent as a property which requires considerably more effort to sell.

The second feature of the real estate brokerage industry is that there are low barriers to working as an agent. Although all states in the U.S. require real estate brokers and salespersons to be licensed, the licensing requirements are generally not perceived as a significant hurdle to becoming a working agent. During the previous period of rapid price appreciation, the national average housing price increased by 83% from 1997 to 2006, making the buying and selling of houses a lucrative business.[6,7] Not surprisingly, the National Association of Realtors witnessed elevated rates of entry, and membership surged from 716,000 to 1,358,000 during that same period.[8] In the subsequent years, housing prices fell slightly, but the total number of housing transactions plunged. Many agents left the market, and the NAR reported a 20% reduction in its membership from 2006

[1] U.S. Department of Housing and Urban Development, Office of Policy Development and Research, "U.S. Housing Market Conditions," 1st, 2nd, 3rd, 4th Quarters 2007, 1st Quarter 2008.

[2] Data source: http://www.federalreserve.gov/releases/z1/current/accessible/b100.htm, B.100 Balance Sheet of Households and Nonprofit Organizations, from Flow of Funds Accounts of the United States published by the Federal Reserve.

[3] While a real-estate broker usually supervises an agent, often as the owner of the firm, and is subject to stricter licensing requirements, we use the terms agent and broker interchangeably.

[4] Data source: Bureau of Economic Analysis, National Income and Product Accounts Table 5.4.5. Private Fixed Investment in Structures by Type. 1929-2008.

[5] See, e.g., Risen (2005).

[6] Data source: http://www.huduser.org/portal/periodicals/ushmc/spring10/hist_data.pdf, Table 10: Repeated Sales House Price Index: 1991-Present. U.S. Department of Housing and Urban Development.

[7] We use the word 'house' to refer to all properties, including condos, single-family, and multi-family dwellings.

[8] Data source: http://www.realtor.org/library/library/fg003, "Field Guide to NAR Membership Statistics, 1908-Present," National Association of Realtors.

to 2009.

As pointed out by Hsieh and Moretti (2003) (hereafter H&M) and others,[9] when real estate agents seek rents resulting from fixed commissions and rising housing prices, their entry behavior may be inefficient because of a business-stealing effect: more agents compete for a fixed supply of properties, with little or no impact on average agent productivity. In support of this argument, H&M document that despite large cross-sectional differences in housing prices across regions, the average earnings of real estate agents are remarkably similar across cities. In addition, cities with high housing prices suffer from low productivity (measured by houses sold per agent) compared to cities with low housing prices. H&M take these findings as evidence that free entry in seeking high commissions is socially inefficient.

In this paper, we build on this earlier work to measure social inefficiency associated with real estate agent entry. We first build a rich data set that contains all properties listed in the Multiple Listing Service (MLS) from January 1998 to December 2007 in all sixty cities and towns within a fifteen-mile radius of downtown Boston. The dataset covers 18,857 agents with 290,378 observations, and includes detailed property characteristics and transaction information for all properties brokered by each agent during both up and down markets.[10] Second, we develop and estimate a dynamic model of agent entry and exit, and conduct a counterfactual analysis in which we examine what would have happened if actual commission on the sale of a house had remained the same in 2007 as it was in 1998.

We focus on two aspects of inefficiency. The first is the amount of income entrants could have earned in some alternative profession had they not worked as agents. This foregone income is socially inefficient because agents' entry does not increase total output in the brokerage industry, but their opportunity wage income is lost as they compete for business with existing agents.[11] The second source of inefficiency is the reduction in service quality (as measured by the average probability of a listed house being sold and the length of time it takes to sell a property), as our empirical evidence shows that inexperienced entrants have worse records on both fronts.

Our main finding suggests that both types of inefficiencies are sizable and hence low barriers to entry have significant costs. Agents' foregone incomes are estimated to vary between 65%-90% of their observed incomes. Listings by new agents are 11% less likely to sell than listings by an agent with six or more years of experience.[12] Moreover, if commissions had remained the same as in 1998, agent entry is reduced by 51%, the exit rate more than doubles, and social savings amount to 30% of industry revenues.

In estimating our dynamic model, we contribute to the literature on the estimation of dynamic discrete choice models by developing a method that makes it feasible to estimate and conduct

[9]Related studies include Crockett (1982), Miceli (1991), Turnbull (1996), and Decloure and Miller (2002).

[10]In the analysis sample, we use 10,088 agents and 257,923 observations; see section 3.2 for details.

[11]The aggregate number of houses sold and bought is largely determined by aggregate housing market conditions and probably does not depend on individual brokers.

[12]While we cannot rule out agent sorting, i.e., experienced agents are good at selecting desirable properties that are easy to sell, our estimation of this effect relies on a large number of controls including detailed physical attributes and zipcode-year fixed effects.

counterfactual analysis using a model with a large number of state variables.[13] Specifically, we estimate the value function non-parametrically using sieves and cast the Bellman equation as a constraint within the MPEC framework. In the appendix, we provide extensive Monte Carlo evidence that the method works well in our application with six state variables.

The remainder of the paper is structured as follows: Section 2 reviews the related literature. Section 3 describes data sources and presents descriptive facts on the real estate brokerage industry. Section 4 develops our econometric model. Section 5 outlines our estimation approach. Section 6 presents and describes our empirical results. Section 7 summarizes our conclusions.

2 Related Literature

This paper is related to the literature on real estate brokers and their impact on the housing market. Among recent empirical studies, H&M study cross-sectional differences in 282 metropolitan areas using data from the 1980 and the 1990 Census of Population and Housing, and provide evidence that free entry by real estate agents is socially inefficient. Using the more recent five percent sample of the 2000 Census of Population and Housing, Han and Hong (2009) examine agents' variable costs of selling houses in a static entry model. Their estimates suggest that free entry leads to a loss of economies of scale: a 10% increase in the number of realtors increases the average variable cost of selling houses by 4.8%. Due to data limitations, they assume all agents are identical and ignore the opportunity cost of entry.

Further afield, there are a number of related papers on real-estate agents. Kadiyali, Prince, and Simon (2009) study dual agency issues in real-estate transactions. Levitt and Syverson (2008) compare home sales by agents who own the property to home sales by agents hired to sell the property. Hendel, Nevo, and Ortalo-Magné (2009) contrast traditional multi-listing services with for-sale-by-owner platforms.

Our paper is also related to the recent econometric models of entry in imperfectly competitive industries (e.g., Collard-Wexler (2008), Dunne, Klimek, Roberts, and Xu (2009), Ryan (2010), and Xu (2008)). While sharing a methodological approach with these papers, our application requires that we differ from these models in a number of ways. Most importantly, the challenges of discretizing a high-dimensional state space and the difficulties of solving equilibria for a model with six to seven state variables lead us to pursue a method that is feasible and easy to implement. Instead of discretizing the state space, we treat it as continuous and approximate the value function using sieves. Our estimator falls into the class of estimators studied by Ai and Chen (2003) and Chen and Pouzo (2009). These papers establish the consistency and the convergence rates of such estimators.

Finally, our paper is related to dynamic models of occupational choice and job matching following Jovanovic (1979). Closely related to our econometric approach, Keane and Wolpin (1994)

[13] Rust (1987) is an early paper in the literature on the estimation of dynamic discrete choice models. Other papers include Ericson and Pakes (1995), Pakes and McGuire (2001), Aguirregabiria and Mira (2007), Bajari, Benkard, and Levin (2007), Pakes, Ostrovsky, and Berry (2007), and Pesendorfer and Schmidt-Dengler (2008).

have a finite-period dynamic model and use backward induction to estimate the value function. In each step, they use fixed point iteration to solve the value function for some state points, and extrapolate to all other state points. Our method may be applicable to similar problems with the advantage that it avoids discretization and allows for the non-parametric estimation of the value function to converge to the true value function as the sample size grows.

3 Data and Descriptive Statistics

3.1 Industry Background

Real-estate agents are licensed experts specializing in real estate transactions. They sell this knowledge about local real estate markets and services associated with the purchase and sale of properties on a commission basis. For home sellers, agents are typically involved in advertising the house, suggesting listing prices, conducting open houses, and negotiating with buyers on behalf of their clients. For home buyers, they search for houses that match their clients' preferences, show the listings, and negotiate with sellers. In addition, they frequently provide suggestions on a host of issues related to changes in property ownership, such as home inspections, obtaining mortgage loans, and finding real estate lawyers.

All states in U.S. require real estate brokers and salespersons to be licensed, but the license requirements are minimal. In Massachusetts, the site of our study, applicants for a salesperson license need to take twenty-four classroom hours of instruction and pass a written exam. The qualification for a broker's license involves additional requirements: one year of residence in Massachusetts, one year of active association with a real estate broker, completion of thirty classroom hours of instruction, passing a written exam, and paying a surety bond of five thousand dollars. Salespersons can perform most of the services provided by a broker, except that salespersons cannot sell or buy properties without the consent of a broker. Both licenses need to be renewed biennially, provided the license holder has received six to twelve hours of continuation education and has paid appropriate fees for each renewal (currently, $93 for a salesperson and $127 for a broker).

In discussions with real estate agents, it appears that these requirements are not perceived as a significant deterrent to working as a realtor. This perception is confirmed by the significant entry from 1998 to 2009, as measured by the total number of members of the National Association of Realtors, reported in Table 1. The number of aggregated home sales increased by 43% from 1998 to 2005, and then sharply declined to only 94% of the 1998 level over the course of four years. The housing price index, as measured by the repeat-sales home price index, peaked in 2007 at 221.1 and dropped to 198.4 by 2009. The number of realtors closely followed the housing price appreciation, and increased rapidly from 2001 to 2006. By 2006, when house prices were highest, the NAR had 1.36 million registered members, 89% more than the membership in 1998. As house prices declined, many agents left the industry, and the membership returned to 1.1 million in 2009.

3.2 Data

The data for this study come from the Multiple Listing Service (MLS) network for the greater Boston area. We collected information on all listed non-rental residential properties in sixty cities and towns within a fifteen-mile radius of downtown Boston, with a total of 18,857 agents and 290,738 observations.[14] For each listed property, we obtained detailed information about the listing (the listing date and price, the listing firm and agent, commissions offered to the buyer's agent, and so on), property characteristics (including address and zip code, the number of bedrooms, bathrooms, and other rooms, the number of garages, age, square footage, lot size, architectural style, whether it has a garden, type of heating, whether it is a condominium, a single family or a multi-family dwelling), the number of days on the market, as well as the sale price and the purchasing agent and firm when a sale occurs.[15] In addition, we merge this data set with the Massachusetts license database to obtain each agent's license history, and append it with the demographic information provided by List Services Corporation. We exclude observations with missing cities or missing listing agents.

Some agents had few transactions and only showed up briefly in our data set. To eliminate agents buying and selling properties for themselves, we keep agents who a) appear in MLS for at least two years, and b) either bought more than one property or list more than 1.5 properties per year. This leaves us with 10,088 agents listing 257,923 properties, about 90% of the sample. The data appendix provides more details on the sample construction.

Our analysis benefits from three sources of variation present in the data: cross-sectional variation among agents (from "green" realtors to established agents with decades of experience), time-series variation in the housing market (from an up market to a down market), and geographical variation in the housing market (the median household income of the most affluent town is three times higher than that of the least wealthy one).

Table 2 reports summary statistics on the number of listed and sold properties, the average sale price, and the number of days it takes to sell a property from the beginning to the end of our sample. The number of listings varied from 20,000 to 23,000 in the late 1990s and early 2000s, but increased to 32,500 in 2005. There was a sharp decline in the number of houses listed in the following years when the housing market suffered from the decline in the aggregate economy. The weakness of the housing market in the latter part of the sample is apparent in the fraction of properties sold: before 2005, 75%-80% of listed properties were sold; in 2007, only 50% were sold. The average sales price was about $350,900 (in 2007 real dollars, adjusted using the urban CPI, series CUUR0100SA0) in 1998 and peaked at $529,200 in 2005. It dropped slightly to $490,000 - $500,000 in 2006 and 2007.

[14] To verify MLS's coverage of all transactions occurred in the cities that we study, we compared Warren Group's changes-of-ownership file based on town deeds records, which we have access to from 1999-2004. This dataset is a comprehensive recording of all changes in property ownership in Massachusetts. The coverage was above 70% for all cities except Boston, which was around 50%. This fact, together with concerns about data quality, lead us to exclude the city of Boston from the empirical analysis.

[15] The number of days on the market is measured by the difference between the listing date and the date the property is removed from the MLS database.

The variation in housing prices is also associated with variation in entry and exit among real estate agents. In Figure 1, we plot the number of active agents, entries, and exits from 1998 to 2007. The number of active agents increased from around 3,800 in 1998 to a peak of more than 5,700 in 2005. This pattern is likely related to house price appreciation during the same time, and parallels the pattern reported in overall market statistics in Table 1. The number of exits was around 400-500 during the early period, but rose to 700-800 in the latter part of our sample when housing market conditions deteriorated. These facts indicate that agent entry and exit are tightly related to aggregate house market trends.

Across agents there is also considerable heterogeneity. The mean number of sold listings per agent reported in Table 2 is about 3.86, and agents working on behalf of buyers brought in roughly the same number of transactions. However, this distribution is highly skewed: both the number of listings sold per agent and the number of houses bought per agent at the 75th percentile is four to six times that of the 25th percentile. Note that the average number of properties sold per listing agent differs slightly from that of an average buying agent because not all agents list and buy in every period. During the down markets of 2006 and 2007, a significant fraction of real estate agents were hit hard: more than 25 percent of agents did not sell any properties at all as the listing agent.

A major factor which contributes to variation in agent performance is agent experience, since experienced agents earn considerably more than new entrants. Figure 2 reports the average annual commissions of agents by the number of years since their entry. The figure reports this calculation for agents who entered in 1999 or later, and excludes agents who were present at the beginning of our sample, because their experience is truncated. Agents who have worked for one year earn on average $20,000. Agents who have worked for two years earn almost double that amount at $35,000. The peak annual commission is just under $50,000 for agents who have six years of experience; however, it is important to note that we have a short panel and cannot follow agents for many years forward if they enter in the later years of our dataset.

In Figure 3, we plot commissions for agents who were present in 1998, breaking the agents into four groups based on their initial earnings. Top quartile agents earned $100,000 or more for most of the years, while the bottom quartile barely received $30,000 in commissions even when housing prices were at their peak. Moreover, agents in the top quartile earn significantly more than than those in the 2nd or 3rd quartile. The earning difference between 2nd and 3rd quartile agents is much smaller and also compresses in down markets. Figure 4 follows these same groups of the 1998 cohort throughout the sample period. It shows that high-commission agents are less likely to exit than low-commission agents and the gap in the fraction of agents staying widens over the years.

4 Model

In this section, we first describe various elements of the model: the state variables, the revenue (or payoff) function, and the transition process of state variables. Our modelling choices are driven by data as well as by computational constraints. Then we present the Bellman equation and the value

function. In the following section, we discuss our solution algorithm and present some Monte-Carlo evidence.

4.1 State variables

We assume that agents' commissions are determined by two sets of (payoff-relevant) variables: aggregate variables and individual characteristics. The aggregate variables include the total number of houses listed on the market, the average housing price, the number of competing agents, as well as the ratio of the number of listings over the total number of properties sold in the previous year. This "inventory-sales ratio" captures the state of supply-demand imbalances in the housing market: it predicts whether a property gets sold, and the amount of time it takes to sell a property. Individual characteristics include an agent's demographics (gender, age), experience, as well as firm affiliation. We assume that the number of houses, the average housing price, and the inventory-sales ratio are exogenously determined by the aggregate supply and demand of the housing market and are not affected by real estate brokers.[16]

Our approximation of the agent's revenue does not include detailed physical attributes for each property that an agent handles. While our data allow us to differentiate large and small houses and various other housing characteristics, including this information as part of the agent's profit function poses several challenges. First, agents do not randomly match with properties. To incorporate housing physical attributes in the revenue function, we need to model how agents and homeowners search for each other. Despite the richness of our data set, it contains no information on home sellers and buyers, and this makes it formidable to model the matching process between households and realtors.

Second, and perhaps more importantly, incorporating a large number of payoff relevant state variables in dynamic models is difficult, especially if the counter-factual analysis requires solving for a new equilibrium. The demand on the richness of the data set to precisely recover the joint transition process of many state variables, the challenges in performing a high-dimensional integral of the unknown value function, as well as rapidly increasing memory requirements, all point to the need to conserve the number of state variables. This is perhaps the major reason that many empirical dynamic models are limited to one or two state variables.

In face of these challenges, we focus on aggregate housing market variables, which abstract away from differences in physical attributes among properties. As shown in Table A.1-Table A.3, the R^2 of the three components of our revenue function varies from 0.32 to 0.86, and the correlation between our model's predicted revenue and the observed revenue is 0.70, even though there are only a limited number of state variables in the revenue function. Since we do not observe commissions paid to listing agents, we assume that the commission rate is 2.5% in this study (i.e., seller's agent and buyer's agent split commissions evenly).

[16] While it is possible that some home owners would not have listed their properties for sale if not for the realtors they know, this is unlikely to account for a significant fraction of total transactions. We ignore such informal channels through which the aggregate housing market is affected by realtors.

4.2 Revenue function

Realtors earn commissions from brokering both the sale and the purchase of homes. We model each type of commission separately.

Agent i's commissions from selling houses depends on his share of houses listed for sale, and the probability that his listings are sold within the contract period. The aggregate variables are the same for all agents in a given market and a given year, so his listing share in year t and market m only depends on his personal characteristics as well as those of his rivals (we omit the market subscript m for notational simplicity):

$$ShL_{it} = \frac{\exp(X_{it}^L \theta^L + \xi_{it}^L)}{\sum_k \exp(X_{kt}^L \theta^L + \xi_{kt}^L)}$$

where X_{it}^L includes agent i's demographics, work experience, firm affiliation, and measures of his skill or reputation (for example, his performance in previous years). ξ_{it}^L is agent i's unobserved quality (observed by all agents, but unobserved by the econometrician), a variable that plays a similar role as the unobserved quality variable in Berry, Levinsohn, and Pakes (1995) and many other discrete choice models. We assume that ξ_{it}^L is independent across periods. Allowing ξ_{it}^L to be correlated over time for agent i makes little difference in the parameter estimate and standard error of θ^L. In addition, the estimated $\hat{\xi}_{it}^L$ exhibits very little persistence over time – the R^2 of regressing $\hat{\xi}_{it}^L$ on its lag is around 0.003, and the AR1 coefficient is only -0.06. Our independence assumption is mainly dictated by the difficulty of incorporating persistent unobserved state variables in dynamic models. There has been some recent progress in this area, see Imai, Jain, and Ching (2009), Norets (2009), and Hu and Shum (2009), but none of them can be directly applied to our application.

The denominator

$$L_t \equiv \sum_k \exp(X_{kt}^L \theta^L + \xi_{kt}),$$

is sometimes called the "inclusive value." It measures the level of competition agent i faces in the brokerage industry. Our implicit assumption is that agents compete in a monopolistic fashion: instead of tracking all rivals' decisions, each agent behaves optimally against the aggregate competition intensity L_t. This assumption is motivated by the fact that there are hundreds of agents per market.

Agents only receive commissions when listings are sold. The probability that agent i's listings are sold is assumed to have the following form:

$$\Pr_{it}^{Sell} = \frac{\exp(X_{it}^S \theta^S)}{1 + \exp(X_{it}^S \theta^S)}$$

where X_{it}^S includes measures of aggregate housing market conditions (total number of houses listed, the inventory-sales ratio that measures market tightness, etc.), as well as his own characteristics.

An agent's total commission from selling listed houses is:

$$R_{it}^{Sell} = r * H_t * P_t * ShL_{it} * \Pr_{it}^{Sell}$$

where 'r' is the commission rate, H_t is the aggregate number of houses listed on the market, and P_t is the average price index.

Analogously, agent i's commission from working as a buyer's agent is:

$$R_{it}^{Buy} = r * H_t^B * P_t * ShB_{it},$$

where H_t^B is the total number of houses bought by all home buyers, P_t is the same as before, and 'ShB_{it}' is his share of the buying market:

$$ShB_{it} = \frac{\exp(X_{it}^B \theta^B + \xi_{it}^B)}{\sum_k \exp(X_{kt}^B \theta^B + \xi_{kt}^B)},$$

where X_{it}^B and ξ_{it}^B are agent i's observed and unobserved characteristics, respectively.

Similar to the listing share, we model

$$B_t \equiv \sum_k \exp(X_{kt}^B \theta^B + \xi_{kt}^B).$$

This is the inclusive value that measures the skills of all buying agents. To reduce the number of state variables, we make the simplifying assumption that $H_t^B = 0.68 H_t$, where 0.68 is the average probability that houses get sold. This assumption is driven by the necessity to economize on state variables and the difficulty of modeling H_t^B as a separate state variable due to its high correlation with H_t.[17] For the same reason, we also group $H_t * P_t$ as HP_t, a single state variable that measures the aggregate size of a housing market.

Combining both types of commissions, agent i's earnings at any given set of payoff-relevant state variables $S_{it} = \{X_{it}^L, X_{it}^S, X_{it}^B, L_t, B_t, HP_t\}$ is:

$$\begin{aligned} R(S_{it}) &= R^{Sell}(S_{it}) + R^{Buy}(S_{it}) \\ &= r * HP_t * (ShL_{it} * \Pr_{it}^{Sell} + 0.68 * ShB_{it}) \end{aligned}$$

4.3 Transition process of state variables

As agents make entry and exit decisions, they care about both their current revenue and their future prospects as realtors. As shown in Table 1, there was a great deal of entry prior to 2005 followed by a plunge in realtor membership after 2006 as housing market conditions deteriorated. Our data are not rich enough to allow us to formulate a model of agent beliefs. Instead, we introduce trend breaks to the transition process before and after 2005.

[17] The correlation between H_t^B and H_t is 0.94 in our sample.

The aggregate state variables are assumed to evolve according to the following AR1 process:

$$S_{t+1} = 1[t <= 2005] * T^1 * S_t + 1[t > 2005] * T^2 * S_t + \eta_t, \tag{1}$$

where $T^i, i = 1, 2$ is a matrix, and η_t is a mean-zero multi-variate normal random variable. We tried to split the sample at year 2005 and estimate a separate transition process for each sub-sample. The R2 for the second part of the sample is very low, and the parameter estimates are much less stable. This should not be surprising since we have only a few periods per market post 2005.[18] An individual agent's characteristics such as his past experience are modeled analogously via an AR1 model.

4.4 The Bellman equation

Agents make career adjustments each period: some continue with their current profession as realtors, others join the real estate brokerage industry (entrants) or leave it (exits). At the beginning of a period, agents observe the exogenous state variables, their own characteristics, as well as two endogenous variables L_{t-1} and B_{t-1} at the end of the previous period.[19] They also observe their private idiosyncratic income shocks and simultaneously make entry and exit decisions. There is no time delay in becoming an agent: they start earning commissions as soon as they become agents and find clients.[20]

Let Z denote exogenous state variables and individual characteristics and Y denote endogenous state variables L and B. The Bellman equation for practicing agent i is:[21]

$$V(Z_{it}, Y_{t-1}) = E_\varepsilon \max_{\varepsilon_{0it}} \left\{ E\left[R(Z_{it}, Y_t)|Z_{it}, Y_{t-1}\right] - c + \varepsilon_{1it} + \delta EV(Z_{i,t+1}, Y_t|Z_{it}, Y_{t-1}) \right. \tag{2}$$

where $E\left[R(Z_{it}, Y_t)|Z_{it}, Y_{t-1}\right]$ denotes agent i's expected commission revenue conditional on observed state variables. Since income shocks are private, agents do not observe rivals' entry and exit when they make their own decisions. Instead, they form an expectation of their commission revenue for the coming period if they continue to work as a broker.

The reservation wage, c, is agent i's opportunity cost – the amount of income he would have earned if he had pursued some alternative profession.[22] It also contains the per-period fixed cost

[18] Another approach to estimate state variables' transition process is to introduce multiple lags and leads as well as high-order polynomials. Given that we have a relatively short data panel, allowing trend breaks is our best attempt to flexibly approximate the structural changes in the aggregate housing market before and after the housing bubble.

[19] L_t and B_t are endogenous because they are determined by all agents' entry and exit decisions jointly. L_t and B_t will increase when more people become realtors and decrease when realtors quit and seek alternative careers.

[20] Many empirical dynamic papers assume that there is a one-period delay in entry. For example, firms pay entry cost at period t, but become an incumbent at period $t+1$. This is a reasonable assumption for firms, because it takes time to install capital and build plants. It is not as appealing here since agents can start earning income as soon as they find clients.

[21] Note that some of the variables (like aggregate state variables) in Z_{it} are common across all agents in a given year.

[22] In lack of a better word, we use "foregone income" and "reservation wage" interchangeably with "opportunity cost".

of being a broker, which includes the expenses of renting office space, the cost of maintaining an active license, the value of the time and energy devoted to building a network. The variable c is the focus of our empirical analysis and is a crucial component of our measure of the social inefficiencies of free entry. Our base specification assumes that c is the same across agents (although we allow it to be different across markets). In our robustness analysis, we model the foregone income as a function of agents' observed attributes.

The model implicitly assumes that "exit" is a terminating action. This is partly driven by data – re-entering agents account for only 9% of our sample – and partly driven by practical considerations. Relaxing this assumption greatly increases the complexity of our estimation and does not add much insight.[23]

Assuming that ε_0 and ε_1 are iid extreme value random variables with standard deviation $\frac{1}{|\beta_1|}$, and denoting the expected payoff as $\bar{R}(Z,Y)$, the original Bellman equation can be rewritten as:

$$\begin{aligned} V(Z_{it},Y_{t-1}) &= E_\varepsilon \max \left\{ \begin{array}{c} \beta_1 \bar{R}(Z_{it},Y_{t-1}) + \beta_2 + \varepsilon_1 + \delta EV(Z_{it+1},Y_t|Z_{it},Y_{t-1}) \\ \varepsilon_0 \end{array} \right. \\ &= \log\left[1 + \exp\left(\beta_1 \bar{R}(Z_{it},Y_{t-1}) + \beta_2 + \delta EV(Z_{it+1},Y_t|Z_{it},Y_{t-1})\right)\right] \end{aligned}$$

where $\beta_2 = -\beta_1 c$.

Individuals who are interested in becoming a realtor can pay a fee (entry cost) and become a broker in the same period. Potential agents enter if the net present value of being an agent is greater than the entry cost κ (up to some random shock). The entry equation is:

$$\begin{aligned} V^E &= E_\varepsilon \max \left\{ \begin{array}{c} -\kappa + \beta_1 \bar{R}(Z,Y) + \beta_2 + \tilde{\varepsilon}_1 + \delta EV(Z',Y'|Z,Y) \\ \tilde{\varepsilon}_0 \end{array} \right. \\ &= \log\left[1 + \exp\left(-\kappa + \beta_1 \bar{R}(Z,Y) + \beta_2 + \delta EV(Z',Y'|Z,Y)\right)\right] \end{aligned}$$

Given our distributional assumptions, the probability that an active realtor quits the brokerage industry is:

$$\Pr(exit_{it}) = \frac{1}{1 + \exp\left(\beta_1 \bar{R}(Z_{it},Y_{t-1}) + \beta_2 + \delta EV(Z_{it+1},Y_t|Z_{it},Y_{t-1})\right)} \tag{3}$$

Let $Y_i = 1$ denote agent i remaining as a licensed broker. The sample likelihood (using only existing agents) is:

$$\begin{aligned} LL(S;\beta) &= \sum_i 1\left[Y_i = 0\right] * \log\left(\Pr(exit_i|\beta)\right) + \\ &\quad \sum_i 1\left[Y_i = 1\right] * \log\left\{1 - \Pr(exit_i|\beta)\right\} \end{aligned}$$

[23] Allowing re-entry requires estimating two value functions. The estimation strategy is similar, except that we need to use the exit choice probability to recast one of the choice-specific value functions as a fixed point of a Bellman equation following Bajari, Chernozhukov, Hong, and Nekipelov (2009).

Provided we are able to solve for EV and calculate choice probability $\Pr(exit_i)$, the structural parameter estimates $\hat{\beta}_1$ and $\hat{\beta}_2$ are chosen to maximize the sample likelihood as stated above.

Analogously, the probability that a potential entry takes place is:

$$\Pr(entry_i) = \frac{\exp\left(-\kappa + \beta_1 \bar{R} + \beta_2 + \delta EV\right)}{1 + \exp\left(-\kappa + \beta_1 \bar{R} + \beta_2 + \delta EV\right)}$$

Let $E_i = 1$ denote agent i entrance. The likelihood of observing N^E new agents with a maximum of $\bar{N}^E$ potential entrants is:

$$LL^E = N^E * \log\left(\Pr(entry_i|\beta)\right) + \left(\bar{N}^E - N^E\right) * \log\left\{1 - \Pr(entry_i|\beta)\right\}$$

We estimate parameters $\hat{\beta}_1$ and $\hat{\beta}_2$ separately from the entry cost κ, which is heavily influenced by the somewhat ad-hoc entry assumption.

4.5 Limitations of our model

We conclude this section by acknowledging some important limitations of our model. First, as mentioned above, the model does not include property attributes and abstracts away from the matching process between agents and home sellers and buyers. In addition, the aggregate housing market is taken as given and not affected by the number of agents. The model does not capture potential benefits consumers derive from agents' marketing behavior (e.g., free pumpkins) because we have no data on these activities. Finally, our model does not allow for serially correlated unobserved state variables, and we assume that agents only keep track of the aggregate intensity of competition (L and B) without following each rival's decision (our monopolistic competition assumption).

5 Solution algorithm: solving for unknown value function $V(S)$

In this section, we describe our solution algorithm. To simplify notation, we omit subscripts throughout this section, and use S to denote the vector of state variables. As explained above, the unknown value function $V(S)$ is implicitly defined by the functional Bellman equation $V(S) = \Gamma(S, E(V(S'|S))$. The ability to quickly compute the value function is a crucial factor in most empirical dynamic models, and in many cases is a determining factor in model specification.

We began our analysis with the traditional approach of discretizing the state space, but met with substantial memory and computational difficulties when we tested our model with four state variables. One of the challenges involves calculating the future value, $EV(S'|S)$, a high-dimensional integral of an unknown function. The quadrature rules we use require evaluating the value function $V(S)$ at quadrature points that do not overlap with grid points. Since $V(S)$ is unknown at any point outside grids, we need to interpolate $V(S)$ from grid points to quadrature points. With four state variables and ten grids each, more than 95% of our computing time was spent on interpolation. As

a result, solving the value function using the Bellman iteration $V^k(S) = \Gamma(S, E(V^{k-1}(S'|S))$ for a given parameter value is painfully slow and often takes a couple of hours. In addition, the memory requirement of discretization increases exponentially, and we ran out of memory on a server with 32GB of RAM when we experimented with five state variables having ten grid points each.

Another factor that discouraged us from discretization is that our data points are far fewer than the size of the state space when the number of state variables is large. Discretizing the state space and solving the value function for the entire state space implies that most of our time is spent solving value function $V(S)$ for states that are never observed in the data (and hence not directly used in the estimation). In addition, as pointed by Bajari, Chernozhukov, Hong, and Nekipelov (2009), with discretization, if in the first stage the transition process is estimated non-parametrically, then it is impossible to have consistent estimates for both the first and the second stage.

We seek an alternative method and approximate the value function $V(S)$ using "sieve estimators," where unknown functions are approximated by parametric functions (the basis functions).[24] This approach has several benefits. First, the sieve approximation eliminates the need to iterate on the Bellman equation to solve the value function, and hence gets rid of the most computationally intensive element of a dynamic analysis. The Bellman equation is instead cast as a constraint of the model that has to be satisfied at the parameter estimates. This drastically reduces the computational burden and makes it feasible to solve for the equilibrium of models with medium to high dimensions. In addition, the algorithm does not waste time calculating the value function in regions of the state space not observed in the sample. It can be a drastic improvement upon methods that require calculating the value function for the entire state space, whose number of elements is often an order of magnitude larger than a typical sample size.[25] The downside of such an approach is that the non-parametric approximation converges to the true value function at a rate slower than the square root of the sample size $\sqrt{N}$. To improve the performance of these nonparametric sieve estimators, Chen and Pouzo (2009) propose a class of penalized sieve minimum distance estimators and show that these estimators achieve the minimax optimal rate for non-parametric mean instrumental variable regression models.

In the following, we present our algorithm in detail, followed by some Monte-Carlo evidence.

5.1 Sieve estimation of the value function

Recall that our Bellman equation is:

$$V(S) = \log\left(1 + \exp\left[\beta_1 \bar{R}(S) + \beta_2 + \delta EV(S'|S)\right]\right). \tag{4}$$

It is a special case of a Hammerstein integral equation studied in Kumar and Sloan (1987), which provides the theoretical foundation for using basis terms to approximate for the value function

[24] See Chen (2007) for a comprehensive survey of sieve estimation in semi-nonparametric models.

[25] For example, with six state variables (which is the number of state variables in our base specification) and ten grids for each, there are 10^6 elements in the state space. This is hundreds of times larger than the number of data points in a typical data set.

$V(S)$.[26] Specifically, let $V(S)$ be approximated by a series of basis functions $u_j(S)$:

$$V(S) \simeq \sum_{j=1}^{K} b_j u_j(S) \tag{5}$$

where the unknown coefficients $\{b_j\}_{j=1}^{K}$ are to be determined. Substitution of equation (5) into (4) leads to a nonlinear equation:

$$\sum_{j=1}^{K} b_j u_j(S) = \log\left(1 + \exp\left[\beta_1 \bar{R}(S) + \beta_2 + \delta \sum_{j=1}^{K} b_j * Eu_j(S'|S)\right]\right)$$

which should hold at all states observed in the data. We choose $\{b_j\}_{j=1}^{K}$ to best-fit this non-linear equation in "least-squared-residuals":

$$\left\{\hat{b}_j\right\}_{j=1}^{K} = \arg\min \left\| \sum_{j=1}^{K} b_j u_j(S_n) - \log\left(1 + \exp\left[\beta_1 \bar{R}(S_n) + \beta_2 + \delta \sum_{j=1}^{K} b_j * Eu_j(S'|S_n)\right]\right)\right\|_2 \tag{6}$$

where $\{S_n\}_{n=1}^{N}$ are state values observed in the data, and $\|\cdot\|_2$ is the $L-2$ norm.[27]

To operationalize this idea, we need to find suitable basis functions $u_j(S)$. There are many possible candidates, including power series, Fourier series, splines, neural networks, etc. In general, the best basis function will be application specific. It is important to use well-chosen basis functions that can approximate the shape of the value function well. A large number of poor basis functions creates various computational problems (non-linear optimization with a large number of unknown parameters) and estimation problems (large bias and variance). A big advantage of our data is that we observe agents' revenue directly. It allows us to exploit information embodied in the revenue function to come up with basis functions that are likely to better approximate the value function. In some cases, economic theory will provide additional shape restrictions on the value function that could improve the performance of the basis function. The following remark is such an example.

Remark 1 *If the revenue function $\bar{R}(S)$ increases (decreases) in S and the transition process TS also increases (decreases) in S, then the value function $V(S)$ increases (decreases) in S.*

The proof is a simple implication of the Contraction Mapping Theorem and appears in the appendix. This property suggests the following strategy: use basis functions (for example, splines) that fit the revenue function $\bar{R}(S)$ to approximate for the value function in the Bellman equation. Since these basis functions are chosen to preserve the shape of $\bar{R}(S)$, they should also capture the shape of the value function.

[26] Kumar and Sloan (1987) show that if the Bellman operator is continuous, and $E(V(S')|S)$ is finite, then sieve approximation approaches the true value function arbitrarily close as the number of sieve terms increases.

[27] There are many ways to formulate the Bellman constraint, and some of them are probably more efficient than others. We leave the choice of the efficient formulation of the Bellman constraint to future research.

Choosing basis terms in high-dimensional models is not a simple matter. To economize on the number of basis terms (which reduces parameter variance and numerical errors), we adopt the 'Multivariate Adaptive Regression Spline' (MARS) method popularized by Friedman (1991) and Friedman (1993) to find spline terms that approximate the revenue function to a desired degree.[28] Once we obtain a set of spline basis terms $\{\hat{u}_j(S)\}_{j=1}^K$ that best fit our revenue function $\bar{R}(S)$, we substitute them for $\{u_j(S)\}_{j=1}^K$ in equation (6). The estimated coefficients $\left\{\hat{b}_j\right\}_{j=1}^K$ are those that minimize the squared difference between the left-hand-side (lhs) and right-hand-side (rhs) of the Bellman equation, where the value function is approximated by $\sum_{j=1}^K \hat{b}_j u_j(S_n)$. As with many other applications of Mathematical Programming with Equilibrium Constraints (MPEC), we impose equation (6) as a constraint and do not explicitly solve for $\left\{\hat{b}_j\right\}_{j=1}^K$ in each iteration of the estimation procedure.[29]

The last technical element in implementing the above method involves integrating the value function. Since $EV(S'|S) = \sum_{j=1}^K b_j * Eu_j(S'|S)$, we can pre-compute the integral of the basis functions $Eu_j(S'|S)$. There are many existing methods for numerical integration. We choose a Quasi-Monte Carlo method because it is easy to implement and one can use the number of simulations to directly control the variance of the numerical integration.[30]

5.2 Monte-Carlo evidence

5.2.1 β_1 and β_2 Estimates in Monte-Carlo simulations

We have conducted extensive Monte-Carlo analyses for our application. In Table B.1, we present parameter estimates from simulating a model identical to the one presented above with four state variables and 2,500 observations.[31] We use four state variables because we want to compare the value function approximated by the sieve estimation with the value function obtained via fixed-point iterations as first proposed in the seminal paper of Rust (1987). The memory requirement of Rust's method becomes computationally prohibitive for models with higher dimensions.

After we solve the value function V^F using Rust's fixed point iteration algorithm, we simulate

[28] MARS repeatedly splits the state space along each dimension, adds spline terms that best improve the fitness according to some criterion function, and stops when the marginal improvement of the fit is below a threshold ($1.0 * 10^{-3}$, for example). We use the R package 'earth' (a package that implements MARS, written by Stephen Milborrow), together with the $L-2$ norm as our criterion function (so that we search for spline knots and spline coefficients that minimize the sum of the square of the difference between the observed revenue and the fitted revenue at each data point).

[29] See Judd and Su (2008) and Dube, Fox, and Su (2009) for illuminating discussions on how to implement MPEC in empirical estimations.

[30] Specifically, we use randomized symmetric Richtmyer points to calculate $Eu_j(S'|S) = \frac{1}{R}\sum_{r=1}^R u_j(S^r|S)$. This is one kind of Quasi-Monte Carlo method that uses carefully selected deterministic sequences of points to increase the integration accuracy, so that the approximation error using N points is on the order of $O(1/N)$, rather than $O(1/\sqrt{N})$ as the standard Monte-Carlo numeric integration. See Bretz and Genz (2009) for more details.

[31] These four state variables are: $HP, inv, L, lagT$. We fix the other two state variables at the sample mean. Details are available upon request.

100 data sets via the following equation:

$$Y_i(S) = 1\left[\beta_1 R(S) + \beta_2 + \delta EV^F(S'|S) + \varepsilon_{i1} > \quad \varepsilon_{i0}\right], i = 1, ..., 2500,$$

where ε_{i1} and ε_{i0} are simulated iid extreme value random numbers. We fix the sample size of each data set at 2,500, which is comparable to the number of observations in our markets.[32] Then we estimate the structural parameters β_1 and β_2 using constrained MLE, where β_1 and β_2 maximize the sample likelihood

$$LL(S; \beta, b_j) = \sum_i 1\left[Y_i = 0\right] * \log\left(\Pr(exit|\beta)\right) + \sum_i 1\left[Y_i = 1\right] * \log(1 - \Pr(exit|\beta)). \tag{7}$$

subject to the constraint that spline coefficients $\{b_j\}_{j=1}^K$ minimize the Bellman violation at β_1 and β_2, as specified in equation (6). We estimate parameter estimates three times, with an increasing number of spline terms going from 9 to 14. The top panel of Table B.1 reports the mean and standard deviation of the parameter estimates for β_1 and β_2.

This exercise is repeated in the bottom panel of Table 1, except that the data $\{Y_i(S)\}$ are generated using the Hammerstein value function V^H. The same spline terms are used for both data generation and estimation; hence, both the β estimates and the spline coefficient estimates are root-N consistent. In the top panel, in contrast, the β estimates are root-N consistent, but the spline coefficients converge more slowly.

The Monte-Carlo evidence validates our approach for our application. In the top panel, the finite-sample bias is around 0.03 when we approximate the value function using nine spline terms, and quickly dropps to 0.01 when the number of spline terms increases to fourteen. The finite sample bias is smaller than the standard deviation of parameter estimates in all cases. The bias in the bottom panel is extremely small: smaller than 0.002 for all cases. This is not surprising since the parameter estimates – a simply application of MLE – are both consistent and efficient.

It is clear that the number of spline terms k is an important component of estimation. The evidence here suggests that k could affect the finite sample bias of parameter estimates (the bias is marginally larger when we increase the spline terms to more than 20). We propose a data driven method to determine k. Let $\left\{\hat{\beta}_1^k, \hat{\beta}_2^k\right\}$ denote the parameter estimates associated with k spline terms. We increase k until parameter estimates converge, when the difference between $\left\{\hat{\beta}_1^k, \hat{\beta}_2^k\right\}$ and $\left\{\hat{\beta}_1^{k-1}, \hat{\beta}_2^{k-1}\right\}$ is smaller than the standard deviation of $\left\{\hat{\beta}_1^k, \hat{\beta}_2^k\right\}$ (which can be estimated using non-parametric bootstrap simulations).[33]

We have estimated many variants of our model. The estimated parameters converge fairly quickly. With three to six state variables and 2,500 observations, the parameters settle down when

[32] Most of our markets have 800 to 2,600 observations, except for Woburn (692), Randolph (700), and Winchester (786).

[33] We also experimented with tighter convergence criteria, for example, parameter estimates "converge" when the difference between $\left\{\hat{\beta}_1^k, \hat{\beta}_2^k\right\}$ and $\left\{\hat{\beta}_1^{k-1}, \hat{\beta}_2^{k-1}\right\}$ is smaller than half the standard deviation of $\left\{\hat{\beta}_1^k, \hat{\beta}_2^k\right\}$.

the number of spline terms increases to 10-15. In general, the bias is small and in most cases insignificant. It is important to note that in scenarios where bias is potentially an issue, one can use various bias reduction techniques proposed in the econometrics literature. Our estimator is fast and easy to compute, and is amenable to most bias reduction techniques that would not have been feasible with most other estimators used in the dynamic discrete choice literature.

6 Empirical analysis

We first discuss our estimates of each element of the model (the revenue function, state variables' transition process), and then report parameter estimates. The subscript m indexes markets.

6.1 Revenue function estimates

Recall that there are three elements in the revenue function: the listing share equation, the buying share equation, and the probability that an agent's listings are sold. We discuss each element in turn.

Agent i's listing share in market m and period t is:

$$ShL_{i,m,t} = \frac{\exp(X^L_{i,m,t}\theta^L + \xi^L_{i,m,t})}{\sum_k \exp(X^L_{k,m,t}\theta^L + \xi^L_{k,m,t})}$$

where $X^L_{i,m,t}$ and $\xi^L_{i,m,t}$ are his observed characteristics and unobserved quality, respectively. Note that there is no constant or aggregate variables in $X^L_{i,m,t}$, because they cancel out from the ratio. Dividing $ShL_{i,m,t}$ by the average share over all agents in market m and period t, and taking logs, we have:

$$\begin{aligned} \ln ShL_{i,m,t} - \ln \overline{ShL_{\cdot,m,t}} &= (X^L_{i,m,t} - \overline{X^L_{\cdot,m,t}})\theta^L + (\xi^L_{i,m,t} - \overline{\xi^L_{\cdot,m,t}}) \\ &= (X^L_{i,m,t} - \overline{X^L_{\cdot,m,t}})\theta^L + \tilde{\xi}^L_{i,m,t} \end{aligned}$$

where $\overline{ShL_{\cdot,m,t}} = \frac{1}{K}\sum_{k=1}^K ShL_{k,m,t}$, $\overline{X^L_{\cdot,m,t}} = \frac{1}{K}\sum_{k=1}^K X^L_{i,m,t}$ and $\overline{\xi^L_{\cdot,m,t}} = \frac{1}{K}\sum_{k=1}^K \xi^L_{i,m,t}$.

We observe an agent's gender, firm affiliation, the number of years as a realtor, as well as the annual number of listings and purchases brokered during our sample period. There is some anecdotal evidence that experience – defined narrowly as the number of years as a realtor – matters, because it takes time to become familiar with local market conditions and individual properties. We examined five aspects of an agent's performance: his number of listings, fraction of listings that are sold, the number of days listings stay on the market, the sale price, as well as the number of purchases. There is a significant difference between agents with fewer than two or three years of experience and agents with five or more years of experience. We follow NAR's convention and define established agents as those with more than five years of experience (exp5).[34]

[34] According to NAR (2007), there is a big difference in annual income between new realtors (with fewer than two years of experience) and established realtors (with six or more years of experience): 65% of new realtors earned less

Agents also differ in skills and the size of their social network. We use agents' total listings and purchases in the previous period (lagT) to approximate for these unobserved attributes. Agents with a lot of past transactions are more likely to gain trust from new customers and to obtain referrals. As shown in the first column of Table A1, lagT is highly predictive: the R2 is as high as 0.45 when it is the sole regressor in the listing share equation. The amount of information contained in this single regressor is surprising considering the amount of heterogeneity we observe among agents. Its coefficient is also economically large: doubling lagT more than doubles the listing share.[35] In contrast, conditioning on past transactions, gender or affiliation with the top three firms (Century 21, Coldwell Banker, and ReMax) does not have much explanatory power. This result is slightly surprising, given the dominant positions of these three firms (they accounted for 44% of realtors and 42% of listings in our sample).

The variable exp5 is significant both statistically and economically: an experienced agent has 15% more listings than a new agent. However, it has a limited explanatory power once lagT is included, partly because it is positively correlated with past transactions. Our preferred specification is column 3, which includes lagT and exp5 as regressors, but we also report results excluding exp5. Some observations with no listings are excluded from this regression because their log-share is not defined. We also exclude 4,000 agents in their first year. Their lagT measure is biased downward because some of them entered the profession in the middle of a calender year. Including these observations does not change the results much, but reduces the R^2 slightly.[36]

One might be concerned that lagT does not fully capture an agent's skill, which can be persistent over time. To address this issue, we regressed the residual estimate $\hat{\xi}^L_{i,m,t}$ on its lags. Interestingly, these residuals exhibit littile persistence over time. The R^2 of the AR1 regression is 0.003, and the AR1 coefficient is -0.06. These results suggest that unobserved persistent attributes are unlikely to be important given our controls. As discussed in section 4.2, in our second-stage estimation, we assume that $\xi_{i,m,t}$ is i.i.d over time and across individuals. Once we have estimated the listing share equation, we obtain our state variable

$$L_{m,t} = \sum_k \exp(X^L_{k,m,t}\hat{\theta}^L + \hat{\xi}^L_{k,m,t})$$

for all markets and periods.

Analogously, we estimate the purchasing share using the following regression:

$$\ln ShB_{i,m,t} - \ln \overline{ShB_{\cdot,m,t}} = (X^B_{i,m,t} - \overline{X^B_{\cdot,m,t}})\theta^B + (\xi^B_{i,m,t} - \overline{\xi^B_{\cdot,m,t}})$$

where $X^B_{i,m,t}$ is the same as $X^L_{i,m,t}$. The results are reported in Table A2. The patterns are very similar to those of Table A1, with the only exception that the coefficient of exp5 is negative. This is driven by the fact that realtors usually begin their careers as buyers' agents, and gradually shift

than $25,000 in 2006, while only 18% of established realtors earned less than $25,000 in 2006.

[35] lagT is standardized to have zero mean and one standard deviation.

[36] These results are available upon request.

to working with sellers as they become more established. Despite the negative coefficient of exp5, its overall effect on revenue is positive: everything else being equal, more experienced agents earn more. As in the listing share equation, we find little persistence in $\hat{\xi}^B_{i,m,t}$. As in the listing equation, we obtain our state variable $B_{m,t} = \sum_k \exp(X^B_{k,m,t}\hat{\theta}^B + \hat{\xi}^B_{k,m,t})$ for all markets and periods.

The third element in the revenue function is the probability that agent i's listings sold:

$$\Pr^{Sell}_{i,m,t} = \frac{\exp(X^S_{i,m,t}\theta^S)}{1+\exp(X^S_{i,m,t}\theta^S)}$$

where $X_{i,m,t}$ includes both aggregate state variables and agent attributes. The results are reported in Table A.3. Besides an agent's skill (captured by lagT), the inventory-sales ratio, which measures the tightness of a market, has a significant and sizable coefficient. Its coefficient varies from -0.12 to -0.08 across different specifications reported in Table A.3 (except for the last column, where the year dummies absorb the variation in the inventory-sales ratio), which means that doubling the inventory-sales ratio will reduce the probability of sales by 30-40%.

The housing markets in the cities we studied experienced a boom and bust in our sample period, with the number of houses sold and housing prices peaking around 2005. The collapse of the housing market triggered the financial crisis and led the economy into recession. In the second half of our sample (2005-2007), houses stayed on the market for a longer period and became much harder to sell. The average fraction of listings that were sold was 0.75 prior to 2005 and plunged to 0.51 afterward. In column 2, we add a trend break to allow different intercepts before and after year 2005. The R2 jumped from 0.14 to 0.86, and the intercepts are statistically different from each other. Once we control for the trend-break intercept, different coefficients for the inventory-sales ratio before and after 2005, gender and firm affiliation, or market and year fixed effects (column 3 to 7) do not seem to matter much.

Once we have estimated payoff parameters, $\theta = \left\{\theta^L, \theta^B, \theta^S\right\}$, we can construct our revenue function as follows:

$$R(S_{i,m,t};\theta) = r * HP_{m,t} * (ShL_{i,m,t} * \Pr^{Sell}_{i,m,t} + 0.68 * ShB_{i,m,t})$$

where we have combined H and P into one state variable HP to measure the aggregate size of the housing market. Since agents do not observe their revenue in the coming period as they make entry and exit decisions (because $L_{m,t}$ and $B_{m,t}$ are determined by agents' decisions simultaneously and are unknown ex ante), we calculate expected revenue using the following (where Z denotes exogenous state variables $HP, inv, lagT$, and Y denotes endogenous state variables L and B):[37]

$$\begin{aligned} &E[R(Z_{i,m,t}, Y_{m,t};\theta)|Z_{i,m,t}, Y_{m,t-1}] \\ &\qquad = \int r * HP_{m,t} * (ShL_{i,m,t} * \Pr^{Sell}_{i,m,t} + 0.68 * ShB_{i,m,t}) dF(Y_{m,t}|Z_{i,m,t}, Y_{m,t-1}). \end{aligned}$$

[37] We also integrate out $\xi^L_{i,m,t}$ and $\xi^B_{i,m,t}$ using their observed distribution.

6.2 State variables' transition

There are four stochastic aggregate state variables: HP, inv, L, B. We standardize the state variables market by market. This is motivated by the considerable size difference across markets: the largest 5 markets have twice as many listings as the smallest 5 markets. Our approach is roughly equivalent to estimating the AR1 regression market by market, but imposing the slope coefficients to be the same.

We report their transition matrix estimates in Table A.5-A.8. Several patterns emerge across all four tables. There is a sizeable level shift before and after 2005. The intercepts are significantly different in almost all specifications for all four state variables. In contrast, allowing different slope coefficients or adding market fixed effects has no observable impact on R2. Year dummies improve the R2 for HP and inv's AR1 regression somewhat, but do not seem to matter for L and B. We exclude year dummies in our analysis. First, the fitness of our preferred specifications without year dummies is reasonably good, with the R2 varying from 0.4 to 0.8. Second, adding year dummies will introduce eight additional state variables and is beyond our current capacity.[38]

We add HP in inv's AR1 regression, because a large number of listings in the previous year tends to increase inventory, leading to a higher inventory-sales ratio. Similarly, HP and inv are added to L and B's AR1 regressions, as both these variables are endogenous and respond to aggregate market conditions.

Finally, we report results for $lagT$ in Table A.4. Unlike the aggregate state variables, there is little change in R2 when we allow for different intercept, different slope coefficients, market dummies or year dummies.

6.3 Structural estimates

Although the housing market is heavily affected by the macro environment and the two move together in general, there is a lot of heterogeneity across markets. In particular, the average commission income in the wealthiest town is about twice that in the poorest town. We estimate our dynamic model separately for each market. Specifically, for each market, we choose $\{\beta_1, \beta_2\}$ and $\{b_j\}_{j=1}^K$ to maximize the following constrained sample likelihood:

$$\begin{aligned} \max_{\beta, b_j} LL(S; \beta, b_j) &= \textstyle\sum_i 1[Y_i = 0] * \log\left[\Pr\left(exit_i|\beta; b_j\right)\right] + \\ \textstyle\sum_i 1[Y_i &= 1] * \log\left[1 - \Pr\left(exit_i|\beta; b_j\right)\right] \end{aligned}$$

$$s.t. \quad \{b_j\}_{j=1}^K = \arg\min \left\| \begin{array}{c} \sum_{j=1}^K b_j u_j(S_n) - \\ \log\left(1+e^{\{\beta_1 \bar{R}(S_n)+\beta_2+\delta \sum_{j=1}^K b_j * Eu_j(S'|S_n)\}}\right) \end{array} \right\|_2$$

where $\Pr\left(exit_i|\beta; b_j\right) = \frac{1}{1+\exp\{\beta_1 \bar{R}(S_n)+\beta_2+\delta \Sigma_j b_j * Eu_j(S'|S_n)\}}$. We use non-parametric bootstraps to estimate standard errors. This estimation is repeated for eight sets of spline basis functions, with an increasing number of terms. We choose the set of parameter estimates $\left\{\hat{\beta}_1^k, \hat{\beta}_2^k\right\}$ whose difference

[38] We have no confidence in our approximation of a value function with twelve state variables.

from the previous iteration $\left\{\hat{\beta}_1^{k-1}, \hat{\beta}_2^{k-1}\right\}$ is smaller than half the size of its standard deviation for both parameters.

We report parameters, their standard errors, the likelihood, the number of observations and the number of spline terms in Table 4. All estimates are significant at a 0.01 level. The estimate of the reservation wage (or opportunity cost) is given by

$$\hat{c}_m = \frac{\beta_{2,m}}{\beta_{1,m}}$$

There is a lot of variation in the level of $\beta_{1,m}$ and $\beta_{2,m}$, but their ratio is much more stable. These estimates suggest that the foregone income is around 65-90% of observed commissions and varies from \$30,000 to \$60,000 for most cities.

6.4 Model's fit

There are several dimensions on which we can compare our model's fit to the observed data. The first is relative to information that we use directly in estimation. In Figure 5, we examine how well our revenue function matches observed commissions each year. We are able to replicate average commissions for most years except for 2005, where the model's prediction is about \$5,000 more than the observed commission (\$65,000 vs. \$70,000). These results are quite decent given that we do not include year dummies in our model.

The other key moment is related to entry and exit of agents. In Figure 6, we examine how well the model predicts the probability of staying each year. The model predicts slightly higher probability of staying in 2000 and 2001 when house prices were rising; in contrast, when house prices fell, the model predicts a lower probability of staying. In Figure 7, we report exit rates by market. Here, the differences between our estimated exit rates and the actual rates are small.

As an out-of-sample test, we compare our estimates to the 2007 median household income in our towns. This involves a validation of our results which does not use information in our dataset. In Figure 8, we plot the estimated reservation wage, sorted from the smallest to the largest, together with the median household income for each town/city in our sample. As the reservation wage increases from the left to the right, the median household also rises, which is reassuring: the foregone income is in general low in poor cities and high in richer towns. There is also a lot of variation in the gap between a reator's reservation wage and a typical household income across towns. In wealthy Boston suburbs, such as Wellesley and Lexington, realtors make considerably less than the median household income. In lower income towns such as Lynn and Revere, a realtor's income is close to the median income in the town.

7 Counter-factual analysis

Our main interest in this section is in computing counterfactuals. This involves developing a method that allows us to solve for a new equilibrium and the associated value function. In our

specific application, we are interested in knowing what would have happened (the market structure, agents' performance, and to some extent, social welfare) if the cost of a real estate transaction has remained the same in the late 2000s as it was in the late 1990s (before the housing boom).

A key element of the counter-factual analysis requires figuring out the new transition process of future L and B conditional on current state variables. Both L' and B' are stochastic because they depend on all agents' entry and exit decisions, which are stochastic. In the estimation, we obtain their transition process from observed data. In the counter-factual analysis, we need to find a new transition process for L' and B' that is "internally consistent."

Consider the thought experiment of realtors facing a changed environment that reduces the payoffs for their services. They first form a belief of the distribution of L' and B' conditioning on today's state variables. Then they "solve" the new Bellman equation and decide individually whether to switch career or not. These stochastic decisions jointly determine the distribution of L' and B'. "Internal consistency" requires the distribution of L' and B' resulting from realtors' optimal behavior to be the same as the belief that factors into their calculation of their future prospects as an agent. In other words, L' is a fixed point of the following Bellman equation (with some abuse of notation):

$$L'(S) = \sum_i 1\left\{\beta_1 R(S_i) + \beta_2 + \delta E_{L',B'}\left[V(S_i')|S_i\right] + \varepsilon_1 > \varepsilon_0\right\} * \exp(X_i^L \theta^L) \tag{8}$$

where S_i is agent i's state (his past experience, etc.), $V(S_i')$ is the equilibrium value function associated with S_i', and $EV_{L',B'}$ is the expectation of V over the distribution of all future state variables, including L', B', and other exogenous state variables. The dependence of EV on the transition process of other state variables is omitted to simplify notation. It is important to note that the 'S' (which is the argument of L') includes all agents' current states.

Given the large number of agents (on average 100-200 per market), and the assumption that their private shocks are iid, the distribution of L' and B' (conditional on current state S) can be approximated by a normal distribution. Here we assume that their new transition process has the same functional form as the one we estimated from data, but with different coefficients:

$$\begin{cases} L' = 1[t <= 2005] * \tilde{T}^1 * S + 1[t > 2005] * \tilde{T}^2 * S + \eta \\ \quad = \qquad\qquad\qquad\qquad\qquad\qquad\qquad \tilde{T}S + \eta \end{cases}$$

where η is a normal random variable. Similarly for the transition process of B'. Under this assumption, "internal consistency" implies that:

$$E(L'|S) = \tilde{T}S = \sum_i \Pr(active_i; \tilde{T}) * \exp(X_i^L \theta^L)$$

where $\Pr(active_i) = 1 - 1/\exp\left(\beta_1 \tilde{R}(S_i) + \beta_2 + \delta \sum_{j=1}^K b_j * E_{L',B'}\left[u_j(S_i')|S_i; \tilde{T}\right]\right)$ for existing agents, and $\Pr(active_i) = 1 - 1/\exp\left(-\kappa + \beta_1 \tilde{R}(S_i) + \beta_2 + \delta \sum_{j=1}^K b_j * E_{L',B'}\left[u_j(S_i')|S_i; \tilde{T}\right]\right)$ for entrants.

$\tilde{T}$ in the middle is equal to $\tilde{T}$ that appears in $\Pr(active_i)$ on the right hand side. The second equation follows from L's definition.

Our approach to solve $\tilde{T}$ can best be described as an iterative one. Starting from $\tilde{T}^0$, we search for $\{b_j\}_{j=1}^K$ such that for a given $\tilde{T}^0$, $\{b_j\}_{j=1}$ minimizes the new Bellman Equation associated with a reduced commission:

$$\{b_j\}_{j=1}^K = \arg\min_b \left\| \sum_{j=1}^K b_j u_j(S_n) - \log\left(1 + \exp\left[\beta_1 \tilde{R}(S_n) + \beta_2 + \delta \sum_{j=1}^K b_j * E_{L',B'}\left[u_j(S_i')|S_i; \tilde{T}^0\right]\right]\right)\right\|_2 \tag{9}$$

Once we obtain $\{b_j\}_{j=1}^K$, we then update the choice probability $\Pr^1(active_i)$ and derive a new estimate of the conditional expectation of $E^1(L'|S)$:

$$E^1(L'|S) = \sum_i \Pr^1(active_i) * \exp(X_i^L \theta^L)$$

Regressing $E^1(L'|S)$ on S delivers $\tilde{T}^1$. We repeat these steps until $\left\|\tilde{T}^r - \tilde{T}^{r-1}\right\|$ is small enough. In practice, these loops are very inefficient. Using the MPEC framework, our counter-factual is once again a constrained optimization:

$$\begin{aligned} & \min \left\| \tilde{T} - \sum_i \Pr(active_i; \tilde{T}) * \exp(X_i^L \theta^L) \setminus S \right\|_2 \\ & s.t. \\ \{b_j\}_{j=1}^K \quad = \quad & \arg\min \left\| \frac{\sum_{j=1}^K b_j u_j(S_n) -}{\log\left(1 + e^{\hat{\beta}_1 \tilde{R}(S_n) + \hat{\beta}_2 + \delta \sum_{j=1}^K b_j * E_{L',B'}\left\{u_j(S'|S_n); \tilde{T}\right\}}\right)} \right\| \end{aligned}$$

where $\sum_i \Pr(active_i; \tilde{T}) * \exp(X_i^L \theta^L) \setminus S$ stands for a least-square projection of $\sum_i \Pr(active_i; \tilde{T}) * \exp(X_i^L \theta^L)$ on S.

Our preliminary results suggest that if real commissions had remained the same over the years as in 1998, entry would decline by 51% on average. In addition, the exit rate would double. Figure 10 plots the counter-factual number of entry, exit and active agents for our sample period. The increase in the number of agents during the housing peak years is much subdued, and the total number of agents would have varied between 1,500 to 2,000 under this new scenario, compared to 4,000 to 5,000 as observed in the data.

8 Conclusion

In this paper we use a new dataset to document stylized facts of entry and exit among realtors in Greater Boston. These facts motivate a dynamic structural model of real estate agent entry which allows us to measure the welfare implications of free entry under a fixed commission structure. The estimates imply a significant change in the market structure under our counterfactual where average commissions remained the same as in 1998: agent entry would be reduced by 51%, the

exit rate would more than double, and the total reduction in the foregone income would amount to about 30% of industry revenues.

Throughout we have focused on measuring the loss of efficiency under the assumption that if there were fewer realtors, overall realtor productivity would not be impacted. In particular, under our assumptions, the social loss is smallest when there is only one agent. Of course, this ignores important aspects of agent heterogeneity and household preferences. If some agents are better than others at selling particular types of properties, a counterfactual with only a small number of agents may neglect this important dimension. Moreover, our efficiency calculation assumes that agents are not capacity constrained. There is evidence that this assumption can be defended for situations like ours. For example, in the cities analyzed by Hsieh and Moretti (2003), the average number of homes sold per agent is more than twice the average for our sample. In our sample, the top quartile agents intermediate four to six times the number of transactions than the lowest quartile. In addition, in the earlier years of our sample before the recent housing price appreciation, agents on average sold and bought 50%-60% more houses than they did in the latter years. All of these empirical patterns suggest that most agents are not capacity constrained, and that reducing the number of agents by half may not have a major impact on service quality.

References

Aguirregabiria, V., and P. Mira (2007): "Sequential Estimation of Dynamic Discrete Games," *Econometrica*, 75, 1–53.

Ai, C., and X. Chen (2003): "Efficient Estimation of Models with Conditional Moment Restrictions Containing Unknown Functions," *Econometrica*, 71, 1795–1843.

Bajari, P., L. Benkard, and J. Levin (2007): "Estimating Dynamic Models of Imperfect Competition," *Econometrica*, 75(5), 1331–1370.

Bajari, P., V. Chernozhukov, H. Hong, and D. Nekipelov (2009): "Nonparametric and Semiparametric Analysis of a Dynamic Discrete Game," Working Paper, Stanford University.

Berry, S., J. Levinsohn, and A. Pakes (1995): "Automobile Prices in Market Equilibrium," *Econometrica*, 63(4), 841–890.

Bretz, F., and A. Genz (2009): *Computation of Multivariate Normal and t Probabilities*. Springer.

Chen, X. (2007): "Large Sample Sieve Estimation of Semi-nonparametric Models, Chapter 76," in *Handbook of Econometrics (Volume 6B)*, ed. by J. J. Heckman, and E. Leamer, vol. 5. Elsevier.

Chen, X., and D. Pouzo (2009): "Estimation of Nonparametric Conditional Moment Models with Possibly Nonsmooth Generalized Residuals," Cowles Foundation Discussion Paper No. 1650R.

Collard-Wexler, A. (2008): "Demand Fluctuations in the Ready-Mix Concrete Industry," Working Paper, New York University.

Crockett, J. H. (1982): "Competition and Efficiency in Transacting: The Case of Real Estate Brokerage," *AREUA Journal*, 10, 209–227.

Decloure, N., and N. G. Miller (2002): "International Residential Real Estate Brokerage Fees and Implications for the U.S. Brokerage Industry," *International Real Estate Review*, 5(1), 12–39.

Dube, J.-P., J. Fox, and C.-L. Su (2009): "Improving the Numerical Performance of BLP Static and Dynamic Discrete Choice Random Coefficients Demand Estimation," Working Paper, Chicago-Booth.

Dunne, T., S. D. Klimek, M. J. Roberts, and D. Y. Xu (2009): "Entry, Exit and the Determinants of Market Structure," Working paper, New York University.

Ericson, R., and A. Pakes (1995): "Markov Perfect Industry Dynamics: A Framework for Empirical Work," *Review of Economic Studies*, 62(1), 53–82.

Friedman, J. H. (1991): "Multivariate Adaptive Regression Splines," *Annals of Statistics*, 19(1).

——— (1993): "Fast MARS," Technical Report, Department of Statistics, Stanford University.

Han, L., and S.-H. Hong (2009): "Testing Cost Inefficiency under Free Entry in the Real Estate Brokerage Industry," University of Toronto, Unpublished mimeo.

Hendel, I., A. Nevo, and F. Ortalo-Magné (2009): "The Relative Performance of Real Estate Marketing Platforms: MLS versus FSBOMadison.com," *American Economic Review*, 5, 1878–1898.

Hsieh, C.-T., and E. Moretti (2003): "Can Free Entry Be Inefficient? Fixed Commission and Social Waste in the Real Estate Industry," *Journal of Political Economy*, 111(5), 1076–1122.

Hu, Y., and M. Shum (2009): "Nonparametric Identification of Dynamic Models with Unobserved State Variables," Caltech, Unpublished mimeo.

Imai, S., N. Jain, and A. Ching (2009): "Bayesian Estimation of Dynamic Discrete Choice Models," *Econometrica*, 77, 1665–1682.

Jovanovic, B. (1979): "Job Matching and the Theory of Turnover," *Journal of Political Economy*, 87, 972–990.

Judd, K. L., and C.-L. Su (2008): "Constrained Optimization Approaches to Estimation of Structural Models," Working paper, Chicago Booth.

Kadiyali, V., J. Prince, and D. Simon (2009): "Is Dual Agency in Real Estate Transactions a Cause for Concern?," Working paper, Indiana University.

Keane, M. P., and K. I. Wolpin (1994): "The Solution and Estimation of Discrete Choice Dynamic Programming Models by Simulation and Interpolation: Monte Carlo Evidence," *Review of Economics and Statistics*, 76, 648–672.

Kumar, S., and I. Sloan (1987): "A New Collocation-Type Method for Hammerstein Integral Equations," *Math. Computation*, 48, 585–593.

Levitt, S., and C. Syverson (2008): "Market Distortions when Agents are Better Informed: The Value of Information in Real Estate Transactions," *Review of Economics and Statistics*, 90, 599–611.

Miceli, T. J. (1991): "The Multiple Listing Service, Commission Splits and Broker Effort," *AREUEA Journal*, 19(4), 584–566.

Norets, A. (2009): "Inference in Dynamic Discrete Choice Models with Serially Correlated Unobserved State Variables," *Econometrica*, 77, 1665–1682.

Pakes, A., and P. McGuire (2001): "Stochastic Algorithms, Symmetric Markov Perfect Equilibria and the 'Curse' of Dimensionality," *Econometrica*, 69(5), 1261–1281.

Pakes, A., M. Ostrovsky, and S. Berry (2007): "Simple Estimators for the Parameters of Discrete Dynamic Games (with Entry-Exit Examples)," *Rand Journal of Economics*, 38(2), 373–399.

Pesendorfer, M., and P. Schmidt-Dengler (2008): "Asymptotic Least Squares Estimators for Dynamic Games," *Review of Economic Studies*, 75(3), 901–928.

Risen, C. (2005): "Realtors vs. the Internet," *The New Republic*, May 2 & 9, 14–15.

Rust, J. (1987): "Optimal Replacement of GMC Bus Engines: An Empirical Model of Harold Zurcher," *Econometrica*, 55(5), 999–1033.

Ryan, S. (2010): "The Costs of Environmental Regulation in a Concentrated Industry," Working paper, MIT.

Turnbull, G. K. (1996): "Real Estate Brokers, Nonprice Competition and the Housing Market," *Real Estate Economics*, 24(3), 293–316.

Xu, D. Y. (2008): "A Structural Empirical Model of R&D, Firm Heterogeneity and Industry Evolution," Working paper, New York University.

Table 1: Real Estate Agents in National Housing Market

Year	Number of Local Realtor Associations	NAR National Membership	Repeat Sales Home Price Index	National Home Sales (1,000s)
1998	1481	718483	124.6	5852
1999	1539	761181	132.0	6063
2000	1524	766560	140.8	6051
2001	1502	803803	150.5	6243
2002	1485	876195	161.1	6605
2003	1471	976960	173.2	7261
2004	1453	1102250	188.2	7981
2005	1445	1265367	205.9	8359
2006	1442	1357732	218.3	7529
2007	1443	1338001	221.1	6428
2008	1437	1197529	208.3	5398
2009	1420	1112645	198.4	5530

Data source: National Association of Realtors; U.S. Department of Housing and Urban Development. URL for HUD data: http://www.huduser.org/portal/periodicals/ushmc/spring10/hist_data.pdf URL for NAR data: http://www.realtor.org/library/library/fg003

Table 2: Number of Properties, Prices, and Days on the Market

	No. of Properties (1000)		Sales Price (2007 $1000)		Days on Market	
Year	Listed	Sold	mean	std. dev	mean	std. dev
1998	23.7	18.3	350.9	295.7	70.4	38.5
1999	22.0	18.1	385.9	320.4	61.5	35.0
2000	20.9	17.2	436.5	367.3	54.4	35.0
2001	22.6	17.6	462.8	365.5	64.5	35.8
2002	23.2	17.9	508.0	375.2	67.7	40.5
2003	25.6	19.4	513.1	362.7	77.5	39.0
2004	28.6	21.4	529.2	363.0	73.7	41.1
2005	32.5	21.1	526.1	355.6	96.8	45.5
2006	31.5	17.2	502.4	361.0	131.9	51.0
2007	27.3	13.6	489.8	364.2	126.2	52.9
All	257.9	181.9	472.1	358.5	85.4	50.0

Data source: Multiple Listing Service. The numbers include all properties listed and sold by 10,088 agents in the Greater Boston Area.

Table 3: Real Estate Agent Listings and Sales

	Num. of	Total Num. of	Num. Sold per Listing Agent			Num. Bought per Buyer's Agent		
Year	Agents	Properties Sold	mean	25th	75th	mean	25th	75th
1998	3,840	18,256	4.75	1	6	3.76	1	5
1999	4,054	18,094	4.46	1	6	4.43	1	6
2000	4,013	17,235	4.29	1	6	4.15	1	6
2001	4,052	17,645	4.35	1	6	3.94	1	6
2002	4,344	17,872	4.11	1	5	3.91	1	5.5
2003	4,791	19,418	4.05	1	5	3.72	1	5
2004	5,328	21,432	4.02	1	5	3.70	1	5
2005	5,763	21,078	3.66	1	5	3.38	1	5
2006	5,671	17,198	3.03	0	4	2.75	1	4
2007	5,227	13,648	2.61	0	3	2.90	1	4
All	10,088	181,876	3.86	1	5	3.61	1	5

Data sourse: Multiple Listing Service for Greater Boston.

Table 4: Parameter Estimates

Market	β_1	std(β_1)	β_2	std(β_2)	Fval	Num. Obs	Num. Spline Terms
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
ARLINGTON	2.55*	(0.58)	-0.75*	(0.19)	245.3	944	9
BROOKLINE	2.35*	(0.33)	-1.03*	(0.16)	278.1	1144	8
CAMBRIDGE	1.38*	(0.23)	-0.62*	(0.12)	467.0	1818	10
CONCORD	1.61*	(0.51)	-0.80*	(0.29)	270.6	879	7
DANVERS	6.77*	(1.32)	-1.45*	(0.29)	285.4	866	18
DEDHAM	3.09*	(0.42)	-1.12*	(0.16)	271.8	934	18
HINGHAM	3.16*	(0.48)	-1.28*	(0.21)	360.7	1208	9
LEXINGTON	3.11*	(0.46)	-1.41*	(0.23)	323.3	1319	9
LYNN	3.74*	(0.43)	-0.96*	(0.13)	719.8	2211	20
MALDEN	2.70*	(0.31)	-0.85*	(0.10)	538.4	1650	11
MARBLEHEAD	3.62*	(0.74)	-1.21*	(0.30)	294.3	1204	10
MEDFORD	2.49*	(0.49)	-0.79*	(0.18)	230.2	823	11
NEEDHAM	2.88*	(0.66)	-1.23*	(0.32)	206.9	858	9
NEWTON	1.90*	(0.31)	-0.89*	(0.16)	540.3	1991	9
PEABODY	3.49*	(0.45)	-0.97*	(0.14)	399.7	1317	10
QUINCY	3.37*	(0.30)	-0.94*	(0.09)	822.6	2569	9
RANDOLPH	3.67*	(0.73)	-0.85*	(0.17)	248.0	700	9
READING	4.04*	(0.63)	-1.01*	(0.19)	331.8	1126	9
REVERE	2.58*	(0.25)	-0.83*	(0.09)	554.1	1787	11
SALEM	3.76*	(0.61)	-0.96*	(0.16)	349.3	1086	11
SOMERVILLE	1.87*	(0.35)	-0.64*	(0.13)	389.5	1080	13
STOUGHTON	4.11*	(0.61)	-1.07*	(0.17)	609.3	1862	13
WAKEFIELD	4.17*	(0.60)	-1.18*	(0.18)	521.0	1694	21
WALPOLE	3.07*	(0.56)	-0.83*	(0.16)	511.1	1458	11
WALTHAM	2.99*	(0.43)	-0.78*	(0.12)	283.3	934	20
WATERTOWN	2.85*	(0.47)	-1.02*	(0.18)	293.6	1123	15
WELLESLEY	1.46*	(0.21)	-0.84*	(0.13)	767.5	2324	19
WEYMOUTH	3.89*	(0.40)	-0.95*	(0.10)	852.6	2504	14
WILMINGTON	4.30*	(0.90)	-1.15*	(0.24)	362.6	965	9
WINCHESTER	2.77*	(0.43)	-1.20*	(0.20)	234.8	786	9
WOBURN	4.97*	(1.26)	-1.17*	(0.31)	248.3	692	10

Note: Parameter standard errors are estimated via 100 bootstrap simulations. Fval is defined as value of the maximized likelihood. * means significant at 1% level.

Figure 1: Entry, Exit, and Number of Active Agents by Year

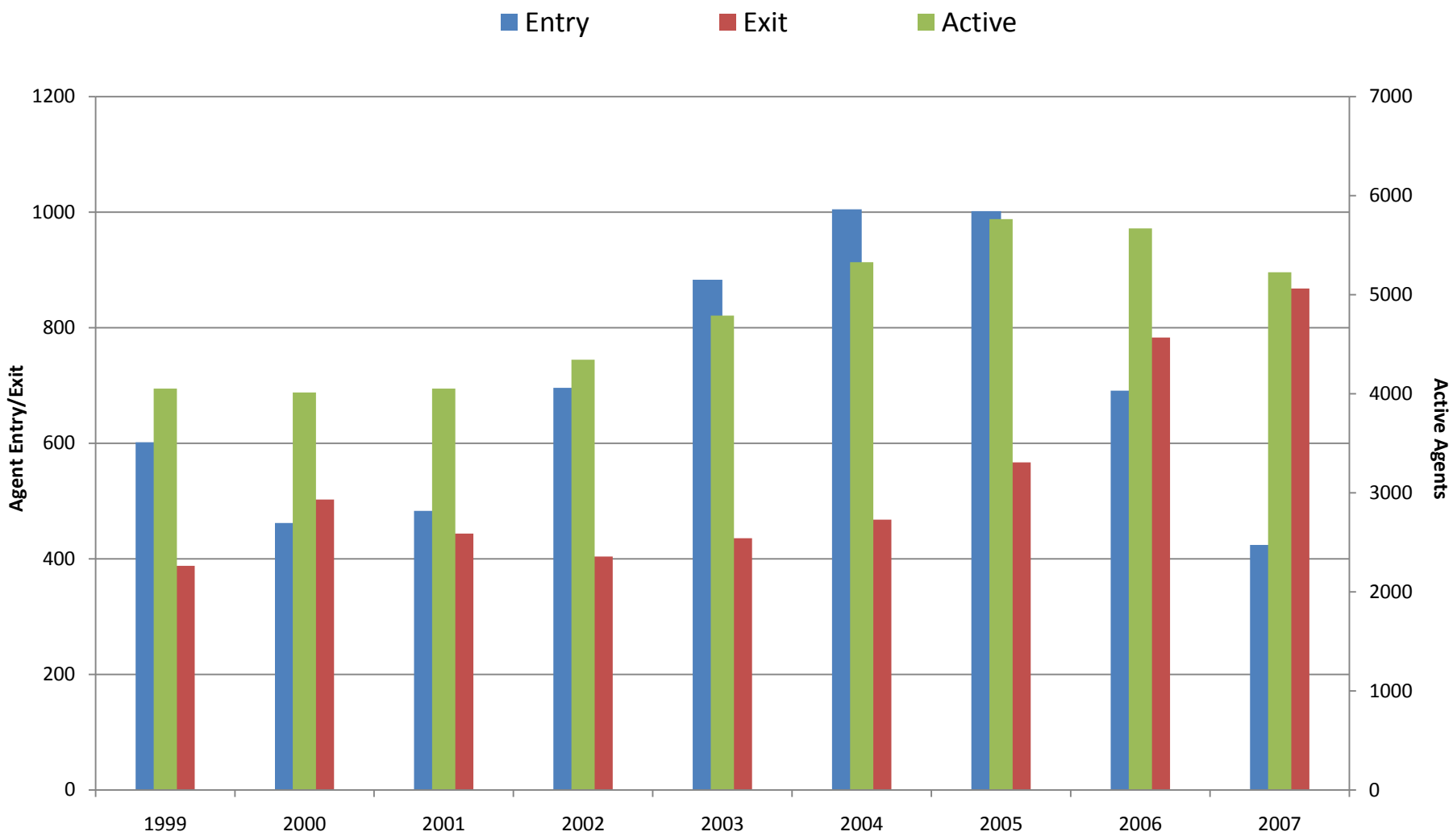

Figure 2: Annual Commissions by Years of Experience for Agents Who Entered in 1999 or After

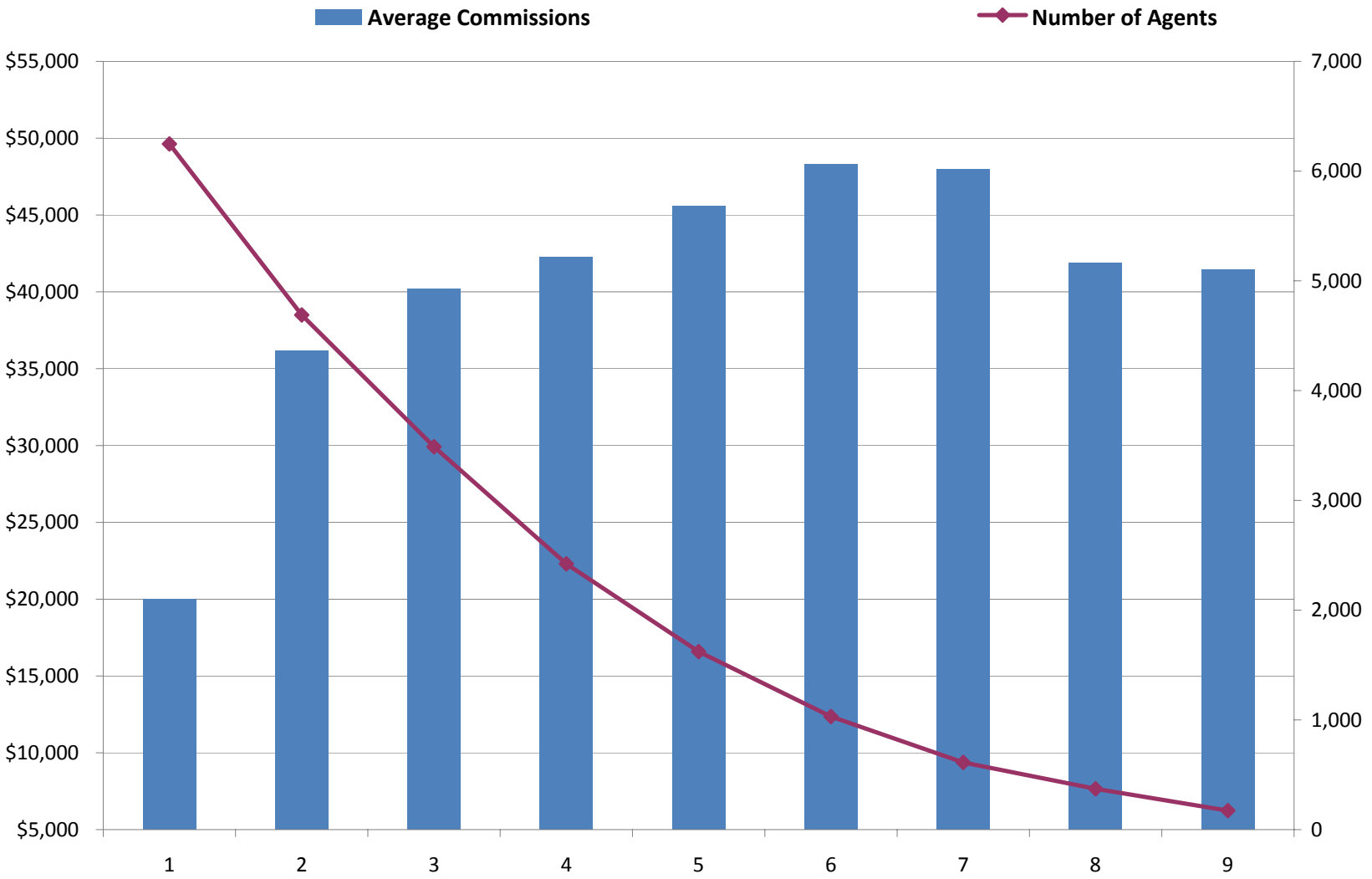

Figure 3: Annual Commissions (2007 $) for the 1998 Cohort by Commission Quartile in 1998

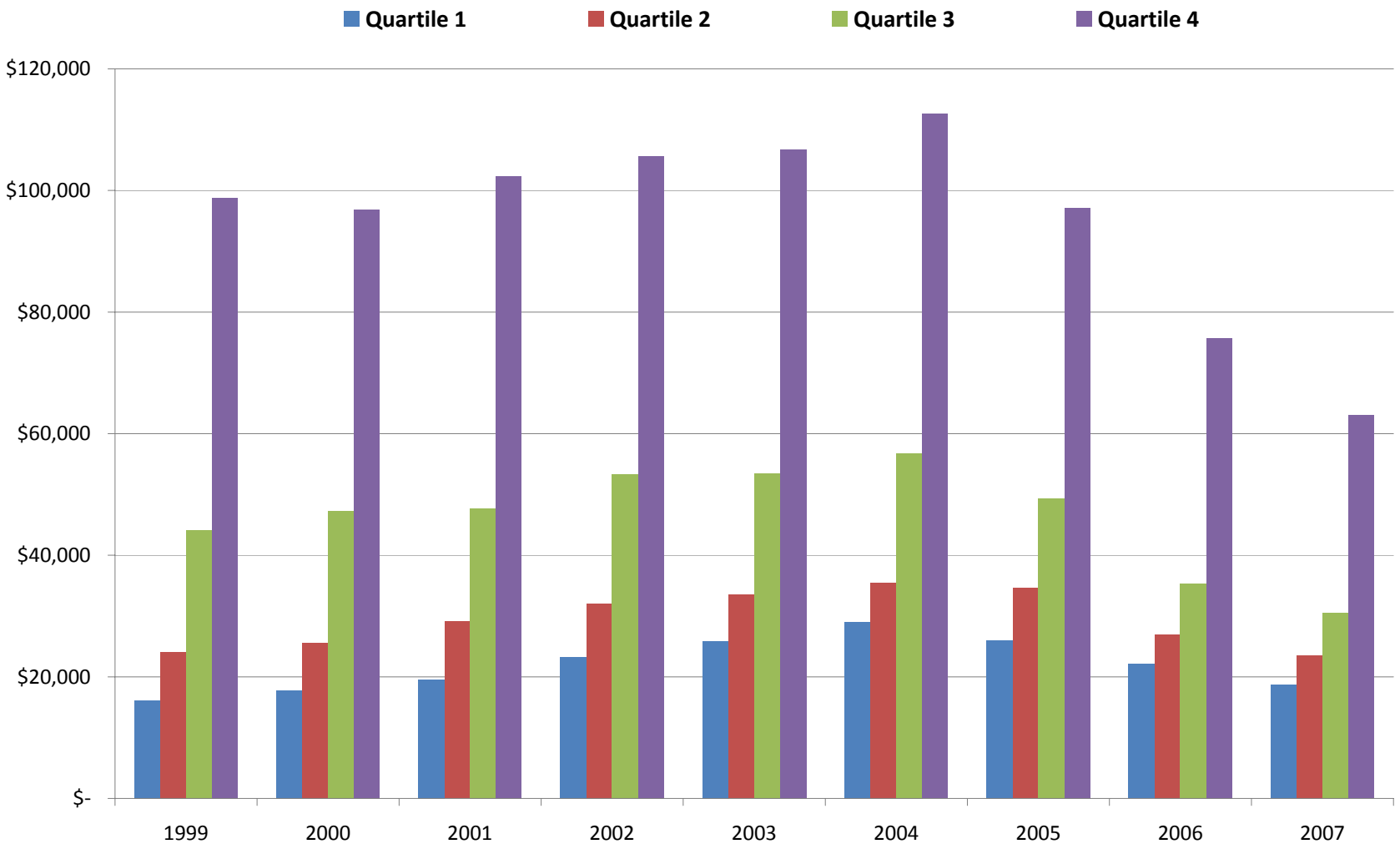

Figure 4: Fraction of Realtors Remaining in 1998 Cohort by Commission Quartile in 1998

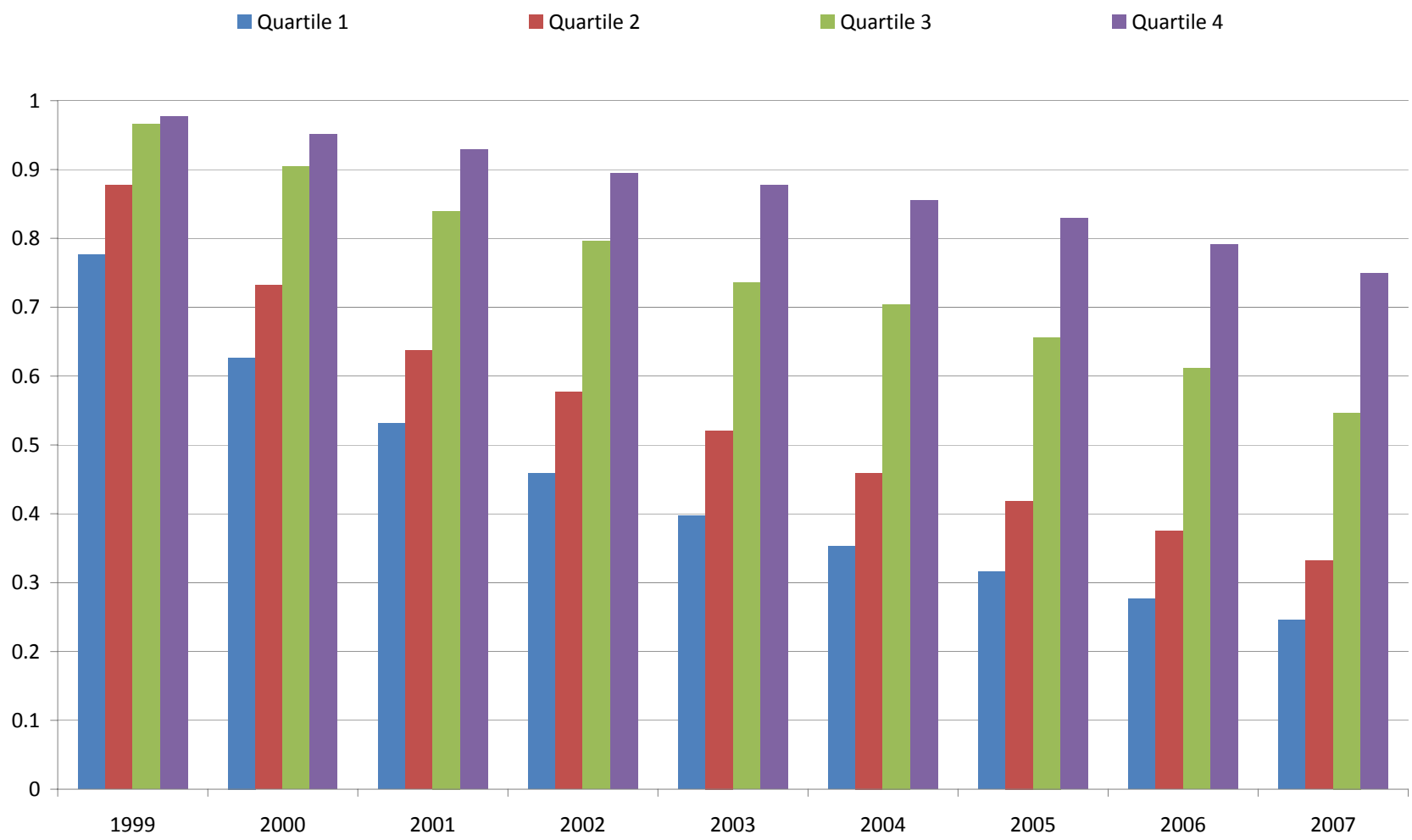

Figure 5: Commissions by Year (2007 $)

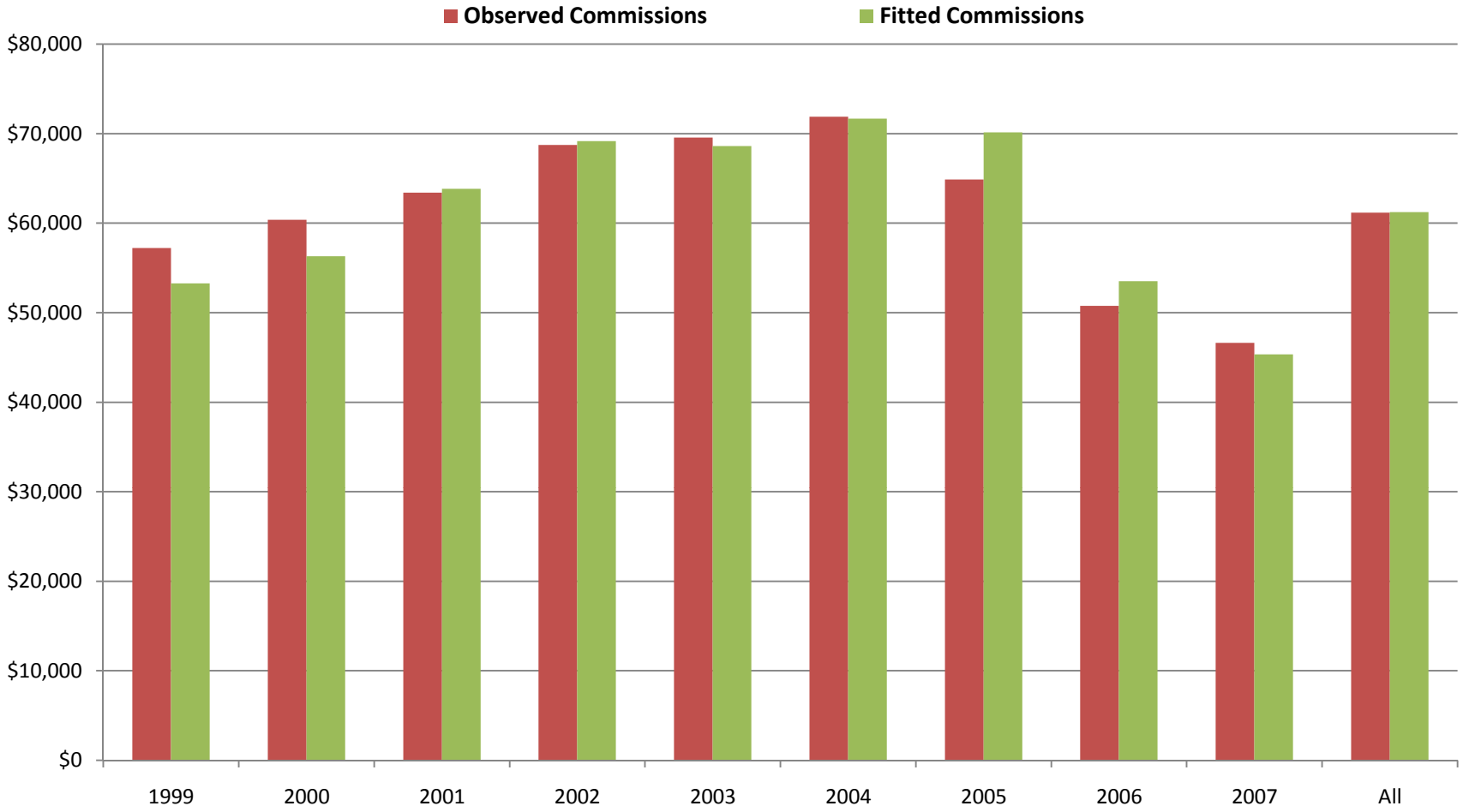

Figure 6: Model's Fit on Probability of Staying by Year

Figure 7: Model's Fit on Probability of Staying by Market

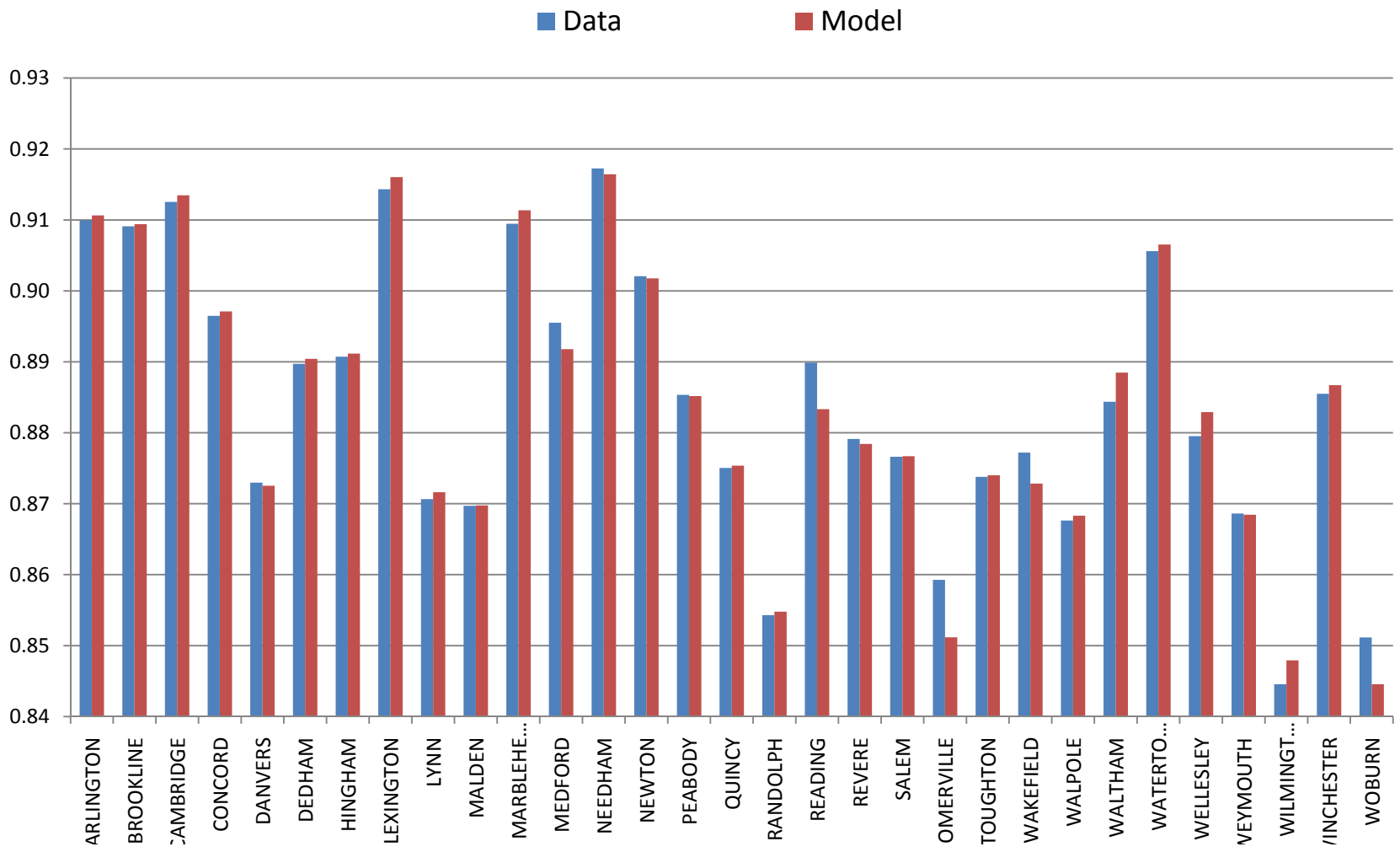

Figure 8: Foregone Income vs. Median Household Income (2007 $)

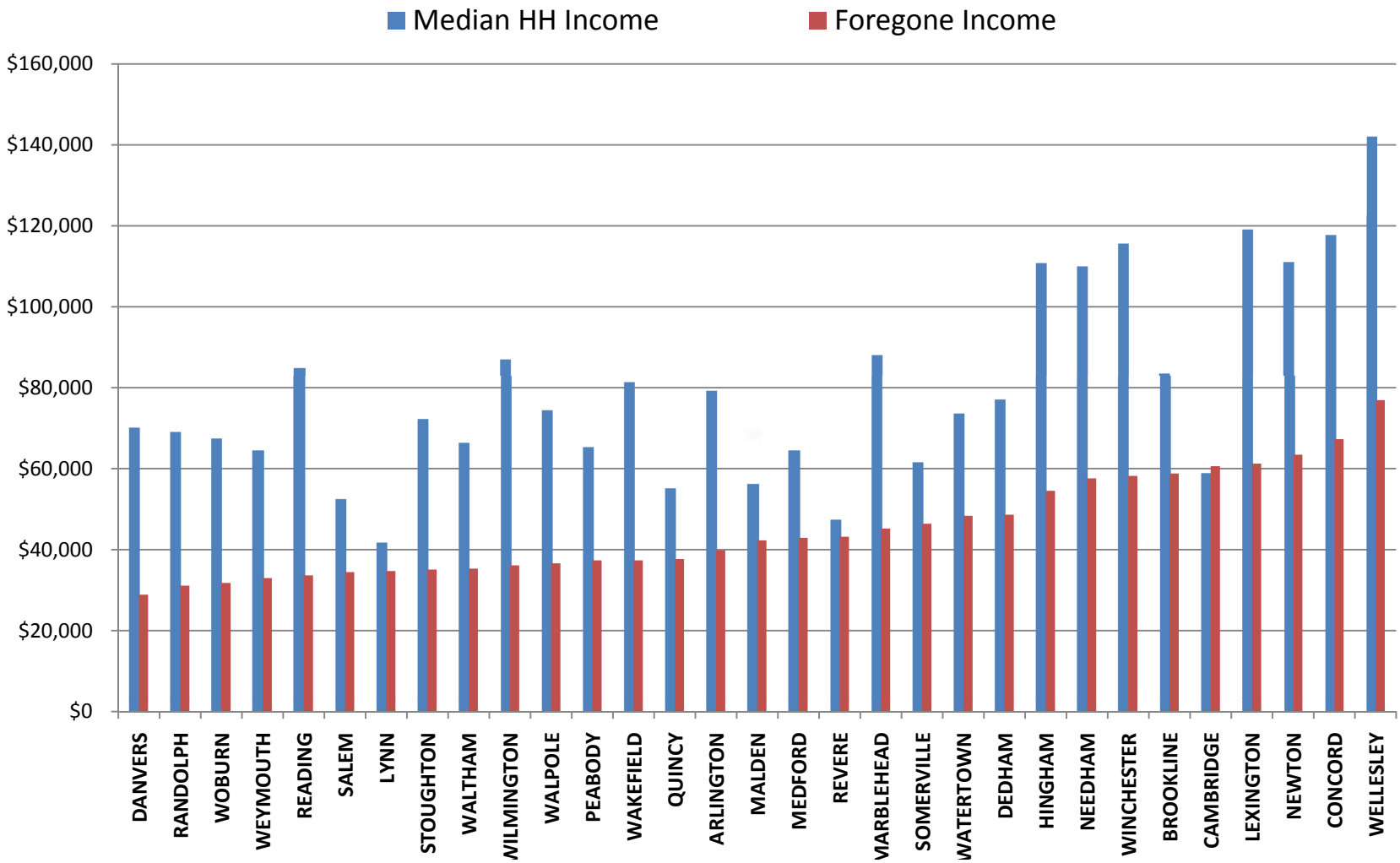

Table A.1: Listing Share Regressions

	(1)	(2)	(3)	(4)
lagT	1.25*	1.25*	1.23*	1.23*
	(0.01)	(0.01)	(0.01)	(0.01)
Male		0.04*		
		(0.01)		
Exp5			0.15*	0.15*
			(0.01)	(0.01)
Century 21				0.01
				(0.01)
Coldwell Banker				-0.04*
				(0.01)
ReMax				0.05*
				(0.01)
Preferred Specification	X			
N	29439	29439	29439	29439
R^2 adjusted	0.4519	0.4522	0.4571	0.4577

Note: '+' stands for p<0.1; '*' stands for p<0.05. 'lagT' is agent i's number of transactions in the previous year. 'Exp5' is one for agents with five or more years of experience.

Table A.2: Buying Share Regressions

	(1)	(2)	(3)	(4)
lagT	0.89*	0.89*	0.91*	0.90*
	(0.01)	(0.01)	(0.01)	(0.01)
Male		-0.05*		
		(0.01)		
Exp5			-0.10*	-0.10*
			(0.01)	(0.01)
Century 21				0.03*
				(0.01)
Coldwell Banker				0.08*
				(0.01)
ReMax				0.11*
				(0.01)
Preferred Specification	X			
N	28316	28316	28316	28316
R^2 adjusted	0.3153	0.3161	0.3188	0.3212

Note: '+' stands for p<0.1; '*' stands for p<0.05. 'lagT' is agent i's number of transactions in the previous year. 'Exp5' is one for agents with five or more years of experience.

Table A.3: Sold Probability Regressions

	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Inv	-0.12*	-0.09*		-0.09*	-0.09*	-0.09*	-0.08*
	(0.00)	(0.00)		(0.00)	(0.00)	(0.00)	(0.00)
lagT	0.03*	0.03*	0.03*	0.03*	0.02*	0.03*	0.04*
	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)
Le05		0.72*	0.72*	0.74*	0.67*	0.71*	0.78*
		(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.01)
G05		0.65*	0.62*	0.67*	0.61*	0.64*	0.71*
		(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)
Inv*Le05			-0.09*				
			(0.00)				
Inv*G05			-0.06*				
			(0.01)				
Male				-0.06*			
				(0.00)			
Exp5					0.07*		0.05*
					(0.00)		(0.00)
Century 21						0.02*	
						(0.00)	
Coldwell Banker						0.05*	
						(0.00)	
ReMax						0.01	
						(0.01)	
Constant	0.71*						
	(0.00)						
Preferred Specification		X					
Market Fixed Effects	No	No	No	No	No	No	Yes
N	29439	29439	29439	29439	29439	29439	29439
R^2 adjusted	0.1402	0.8631	0.8632	0.8642	0.8648	0.8637	0.8699

Note: '+' stands for p<0.1; '*' stands for p<0.05. 'Inv' is the sales-inventory ratio in the previous year; 'Le05' and 'G05' are indicators for year<=2005 and year>2005, respectively. See Table A.2 for the explanation of lagT and Exp5.

Table A.4: Experience (LagT) Regressions

	(1)	(2)	(3)	(4)	(5)	(6)	(7)
lag_lagT	0.76*	0.76*	0.76*		0.76*		0.75*
	(0.00)	(0.00)	(0.00)		(0.00)		(0.00)
Exp5		0.02*			0.02*	0.02*	0.02*
		(0.00)			(0.00)	(0.00)	(0.01)
Le05			0.01*	0.01*	-0.01	-0.01	0.02
			(0.00)	(0.00)	(0.00)	(0.00)	(0.02)
G05			-0.04*	-0.03*	-0.05*	-0.05*	-0.02
			(0.00)	(0.00)	(0.01)	(0.01)	(0.02)
lag_lagT*Le05				0.77*		0.77*	
				(0.00)		(0.00)	
lag_lagT*G05				0.75*		0.75*	
				(0.01)		(0.01)	
Constant	-0.01*	-0.02*					
	(0.00)	(0.00)					
Preferred Specification			X				
Market Fixed Effects	No	No	No	No	No	No	Yes
N	27619	27619	27619	27619	27619	27619	27619
R^2 adjusted	0.5707	0.571	0.5834	0.5834	0.5835	0.5836	0.5853

Note: '+' stands for p<0.1; '*' stands for p<0.05. See Table A.3 for variable definitions.

Table A.5: Market Level Housing Value Regression

	(1)	(2)	(3)	(4)
lag_HP	0.63*	0.98*		0.99*
	(0.04)	(0.04)		(0.04)
Le05		0.46*	0.46*	0.54*
		(0.04)	(0.04)	(0.17)
G05		-0.75*	-0.67*	-0.68*
		(0.07)	(0.17)	(0.18)
lag_HP*Le05			0.99*	
			(0.04)	
lag_HP*G05			0.91*	
			(0.14)	
Constant	0.16*			
	(0.04)			
Preferred Specification		X		
Market Fixed Effects	No	No	No	Yes
N	248	248	248	248
R^2 adjusted	0.522	0.7366	0.7358	0.7051

Note: '+' stands for p<0.1; '*' stands for p<0.05. See Table A.3 for variable definitions.

Table A.6: Inventory Regression

	(1)	(2)	(3)	(4)	(5)
lag_Inv	0.92*	0.45*	0.04		0.04
	(0.05)	(0.06)	(0.05)		(0.05)
lag_HP		0.53*	0.37*	0.41*	0.38*
		(0.05)	(0.04)	(0.04)	(0.04)
Le05			-0.24*	-0.28*	-0.30*
			(0.04)	(0.05)	(0.15)
G05			1.08*	0.97*	1.01*
			(0.07)	(0.09)	(0.16)
lag_Inv*Le05				-0.07	
				(0.09)	
lag_Inv*G05				0.11	
				(0.07)	
Constant	0.28*	0.18*			
	(0.04)	(0.04)			
Preferred Specification			X		
Market Fixed Effects	No	No	No	No	Yes
N	248	248	248	248	248
R^2 adjusted	0.5516	0.7092	0.8473	0.8483	0.8275

Note: '+' stands for p<0.1; '*' stands for p<0.05.

Table A.7: Inclusive Value Regression for Listing Share

	(1)	(2)	(3)	(4)	(5)	(6)
lag_L	0.46*	0.34*	0.28*	0.21*		0.23*
	(0.05)	(0.05)	(0.07)	(0.06)		(0.06)
HP		0.54*		0.81*	0.81*	0.80*
		(0.06)		(0.07)	(0.07)	(0.07)
Inv			0.27*	-0.75*	-0.75*	-0.75*
			(0.08)	(0.12)	(0.12)	(0.13)
Le05				-0.43*	-0.44*	-0.3
				(0.07)	(0.08)	(0.25)
G05				1.45*	1.44*	1.56*
				(0.19)	(0.22)	(0.32)
lag_L*Le05					0.21*	
					(0.07)	
lag_L*G05					0.23+	
					(0.13)	
Constant	0.12*	0.02	0.09			
	(0.05)	(0.05)	(0.05)			
Preferred Specification				X		
Market Fixed Effects	No	No	No	No	No	Yes
N	248	248	248	248	248	248
R^2 adjusted	0.2232	0.4307	0.2551	0.5425	0.5406	0.4952

Note: '+' stands for p<0.1; '*' stands for p<0.05.

Table A.8: Inclusive Value Regression for Buying Agent Share

	(1)	(2)	(3)	(4)	(5)	(6)
lag_B	0.44*	0.32*	0.36*	0.28*		0.30*
	(0.05)	(0.05)	(0.07)	(0.06)		(0.06)
HP		0.49*		0.70*	0.72*	0.70*
		(0.06)		(0.07)	(0.07)	(0.08)
Inv			0.15*	-0.58*	-0.59*	-0.58*
			(0.07)	(0.13)	(0.13)	(0.14)
Le05				-0.28*	-0.30*	-0.29
				(0.08)	(0.08)	(0.28)
G05				1.00*	0.81*	0.97*
				(0.21)	(0.24)	(0.35)
lag_B*Le05					0.23*	
					(0.07)	
lag_B*G05					0.47*	
					(0.13)	
Constant	0.11*	0.02	0.09+			
	(0.05)	(0.05)	(0.05)			
Preferred Specification				X		
Market Fixed Effects	No	No	No	No	No	Yes
N	248	248	248	248	248	248
R^2 adjusted	0.2119	0.3863	0.2229	0.4384	0.4429	0.3818

Note: '+' stands for p<0.1; '*' stands for p<0.05.

Table B.1: Parameter Estimates Using Four State Variables

		1st Set		2nd Set		3rd Set	
Simulate Data w/ V_F	β_0	Mean	Std	Mean	Std	Mean	Std
β_1	1	1.03	0.06	1.01	0.06	1.01	0.06
β_2	-1	-1.03	0.05	-1.02	0.05	-1.01	0.05
Num of Basis Terms		9		11		14	
Simulate Data w/ V_H	β_0	Mean	Std	Mean	Std	Mean	Std
β_1	1	1.00	0.06	1.00	0.06	1.00	0.06
β_2	-1	-1.00	0.05	-1.00	0.05	-1.00	0.05
Num of Basis Terms		9		11		14	

Note: the revenue function R(S)=(7.3*S1+34.2)*{exp(0.1+1.2*S4)/(215+28*S3)*exp(1.4-0.5*S2+0.2*S4)/(1+exp(1.4-0.5*S2+0.2*S4))+exp(0.9*S4-0.1)/154)}. 100 Monte-Carlo simulations.