Supplemental Material to “The Costs of Free Entry: An Empirical Study of Real Estate Agents in Greater Boston”

Panle Jia Barwick* Parag A. Pathak†

First version: August 2011
This version: September 2012

Contents

1 Data **2**
1.1 Data sources . . . 2
1.2 Market definition and the Timing Convention . . . 2
1.3 Market-level state variables . . . 2
1.4 Agent characteristics . . . 3

2 Empirical Analysis **4**
2.1 Revenue function . . . 4
2.2 State variables' transition . . . 5
2.3 Estimates of β . . . 6

3 Counterfactual Analyses **6**

4 Evaluating Value Function Approximation **8**
4.1 Value function approximation . . . 8
4.2 Value function comparison . . . 9
4.3 Parameter estimates in Monte Carlo simulations . . . 10

This paper contains supplemental material to Barwick and Pathak (2012), which we refer to hereafter as BP1. The contents are summarized as follows. Section 1 describes how we classify markets and construct variables used in BP1. Section 2 presents alternative specifications of the revenue function and state variable transition process not reported in BP1. Section 3 provides additional details of the counterfactual analyses. Section 4 compares the value function approximation with two other approaches commonly used in the dynamic discrete-choice literature, and presents Monte-Carlo evidence of the magnitude of parameter bias generated by value function approximation in our application.

*Department of Economics, MIT and NBER. Email: pjia@mit.edu. Barwick gratefully acknowledges the research support of the National Science Foundation via grant SES-1061970.
†Department of Economics, MIT and NBER. Email: ppathak@mit.edu.

1 Data

This section briefly explains how we define a market and calculate state variables. Table A1 presents the summary statistics for all variables used in the estimation.

1.1 Data sources

The main data source is the MLS Property Information Network (MLSPIN) from New England. We collected all available information on the MLSPIN system for each property listing in any town within a 15 mile radius of downtown Boston (City Hall) through December 2007. The data include detailed property characteristics for each listed house such as the number of bedrooms, bathrooms, and lot size, the name and ID of the listing agent and listing office, and the initial list date and price. If a property is sold, then the name and ID of the buying agent and the buying office as well as the sale date and price are recorded. It appears that the MLS does not have complete coverage of transactions for early years of the dataset, so we begin our sample in 1998 and keep all properties that are listed and sold through December 2007. We omit properties that are active listings at the end of our sample.

1.2 Market definition and the Timing Convention

There are 59 cities in the MLS dataset (excluding Boston). Twenty-eight cities are small with fewer than 600 agent-year observations. We group them into contiguous markets as follows: Avon is grouped with Stoughton; Bedford with Lexington; Belmont with Watertown; Braintree with Weymouth; Burlington with Wilmington; Canton with Stoughton; Chelsea with Revere, Cohasset with Hingham; Dover with Wellesley; Everett with Malden; Holbrook with Randolph; Hull with Weymouth; Lincoln with Concord; Lynnfiled with Peabody; Medfield with Walpole; Melrose with Wakefield; Milton with Quincy; Nahant with Swampscott; North Reading with Reading; Norwood with Walpole; Saugus with Lynn; Sharon with Stoughton; Stoneham with Wakefield; Swampscott with Marblehead; Wayland with Wellesley; Weston with Wellesley; Westwood with Dedham; and Winthrop with Revere. The rest of the markets are the same as an MLS city. We group cities that are adjacent, with similar median household income, and have significant amount of cross-city listings by agents.

An agent is an entrant in year t if he is not observed in our sample in or before year $t-1$, but is observed in year t. An agent is active in year t if he is observed in year t. An agent is an incumbent in year t if he is active in year t and entered in previous years. An agent exits in year t if he is observed in year $t-1$, but not in year t.

1.3 Market-level state variables

The aggregate number of listings H_{mt} in market m and year t is the total number of listings between January and December of year t in market m. The average housing price index P_{mt} is an equally-weighted average of the sold price over all listings sold in market m and year t. (Results are nearly identical if we average over all sold listings that were *listed* in year t in market m.) HP_{mt} is the product of H_{mt} and P_{mt}.

To calculate the market tightness measure, the "inventory-sales ratio" (denoted Inv), we proceed as follows. For each month/year/market, we calculate the total number of new listings, the total number of houses taken off the market in this month (which includes sold properties, unsold properties that are withdrawn or whose contract expires in that month). Then we calculate the cumulated sum of listings for each month/year/market, as well as the cumulated sum of properties taken off the market for each month/year/market. Inventory is cumulated listings minus cumulated properties-off-market in the previous month. For example, the inventory is calculated as follows:

Month	Listing	Sold	Inventory
1	100	50	100
2	80	60	100+80-50
3	70	40	100+80+70-50-60
...			

We take the ratio of the current month inventory to total number of listings sold in the previous year in market m, and average over twelve months to derive the annual inventory-sales ratio for each market in each year.

1.4 Agent characteristics

An agent's experience is measured by his number of years as a broker. This variable is generated by merging our MLS data with Massachusetts' license data. The Massachusetts Division of Professional Licensure maintains a database of all real-estate-broker and salesperson licenses issued in the past. We match these records with our MLS data by agent names.

For an agent who first appears in our sample in 1999 or after, we use the year he is first observed as the year he becomes an agent. This accounts for 6,248 agents, or 62% of our sample. For the rest of 3,840 agents who first appear in our sample in 1998 or before, we use their license year if their name is unique in both MLS and the license database. We are able to locate a license record for 2,644 of these agents, a match rate of 69%. For the remaining 1,196 agents (12% of the sample) whom we could not locate in the license database or whose names are not unique, we use the year they are first observed in MLS as the year they become an agent.

Information on agents' gender is provided by List Services Corporation, which links names to gender based on historical census tabulations. For the few agents who work at multiple offices in any given year, we use the office with which he conducted the largest number of transactions as his office affiliation.

We measure an agent skill by his total number of transactions in the previous year. This variable, s_{it}, is highly skewed and varies between 0 to 87, with a median of 6. We truncate it at 20, which is the 90% percentile of the distribution, and normalize it to have zero mean and 0.5 standard deviation to prevents overflowing or underflowing errors.

2 Empirical Analysis

In this section, we present alternative specifications of the revenue function and state variable transition process not reported in BP1, as well as the structural parameters $\{\beta_1, \beta_{2m}\}_{m=1}^{31}$ associated with different sets of spline functions.

2.1 Revenue function

The first element of the revenue function is the listing share equation:

$$\ln ShL_{imt} - \overline{\ln ShL_{\cdot mt}} = (X_{imt}^L - \overline{X_{\cdot mt}^L})\theta^L + (\xi_{imt}^L - \overline{\xi_{\cdot mt}^L}) = (X_{imt}^L - \overline{X_{\cdot mt}^L})\theta^L + \tilde{\xi}_{imt}^L, \tag{1}$$

where X_{imt} includes gender, firm affiliation, experience, and skill. About 11% of observations have zero listing in some years, and the log of their listing shares is not defined. In addition, s_{it} is biased downward for entrants and the second-year agents who become a broker in the middle of a calendar year during their first year. These observations are excluded in regressions reported in Table A2, leaving us a total of 32,237 agent-year observations. Including these observations only slightly reduces s_{it} coefficient.

As discussed in BP1, s_{it} is an important predictor of listing shares. Conditioning on s_{it}, gender or affiliation with the top three firms (Century 21, Coldwell Banker, and ReMax) leads to no improvement in R^2, even though these firms account for more than 40% of listings. Following the convention of the NAR (2007), we define established agents as those with six or more years of experience ($exp6$). The variable $exp6$ is statistically and economically significant: an experienced agent has 14% more listings than an inexperienced one. However, it has a limited explanatory power once s_{it} is included, as might be expected since it is highly correlated with s_{it}.

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

$$\ln ShB_{imt} - \ln \overline{ShB_{\cdot mt}} = (X_{imt}^B - \overline{X_{\cdot mt}^B})\theta^B + \tilde{\xi}_{imt}^B,$$

where X_{imt}^B is the same as X_{imt}^L. The results, reported in Table A3, have patterns that resemble those in Table A2, except that the R^2 is slightly lower: 0.3 vs. 0.44 in Table A2. In addition, coefficients of $exp6$ and $male$ are negative. The negative coefficient of $exp6$ reflects the fact that realtors usually begin their careers as buyers' agents, and gradually shift to working with sellers after they become more established, a phenomenon that we documented in Table 3 in BP1. As with $\hat{\tilde{\xi}}_{imt}^L$ in the listing share equation discussed in BP1, we find little persistence in $\hat{\tilde{\xi}}_{imt}^B$. The Arellano-Bond estimate of the autoregressive coefficient is -0.04.

The third element in the revenue function is the probability that agent i with a total of L_{imt} listings sells T_{imt} of them:

$$\Pr(T_{imt}|L_{imt}) = \binom{L_{imt}}{T_{imt}} \Pr(\text{sell}_{imt})^{T_{imt}} \left(1 - \Pr(\text{sell}_{imt})\right)^{L_{imt}-T_{imt}}.$$

We report MLE estimates of θ^S in Table A4. All columns share two common regressors: the inventory-sales ratio Inv_{mt} and skill s_{it}. Variable Inv_{mt} has a significant and sizable coefficient that varies from -0.52 to -0.35 across different specifications. The coefficients of s_{it} are also highly significant and varies from 0.10 to 0.21. A standard deviation increase in the inventory-sales ratio reduces the probability of sales by 11-16%, while a standard deviation increase in s_{it} improves the probability of sales by 3-6.5%.

Houses stay on the market for a longer period and become much harder to sell in the second half of our sample (2005-2007). The average sales probability is 0.75 prior to 2005 and plunges to 0.51 afterward. In columns (2)-(6) of Table A4, we add a trend break to allow for different intercepts before and after year 2005. These intercepts are statistically different from each other in all cases. In columns (3)-(6), we experiment with gender, experience, firm affiliation, and market fixed effects. Interestingly, conditioning on skills s_{it}, female brokers are more likely to strike a deal: they sell 6% more of their listings than their male counterpart. Agents with 6 or more years of experience also exhibit higher performance and are 8.7% more likely to carry through a transaction than inexperienced agents. Brokers associated with Coldwell Banker appear to do slightly better than those affiliated with Century 21, ReMax, or other firms. However, conditioning on Inv_{mt}, s_{it}, and the trend-break, gender, experience, or firm affiliation does not improve the model's fit. Our preferred specification is column (6), which also include market fixed effects to control for time-invariant market-level housing conditions that affects whether a property gets sold.

2.2 State variables' transition

There are five stochastic state variables: agent skill measure s and aggregate state variables HP, Inv, L, and B. Estimates of AR(1) models for agent skills are reported in Table A5. The autoregressive coefficients are similar across specifications and vary between 0.74 and 0.75. As in the revenue regressions, agent gender and firm affiliation have little impact on R^2. There is a noticeable increase in R^2 when the constant term is allowed to differ before and after 2005. Our preferred specification is column (5), where the regressors include the lag of skill as well as trend-break dummies.

Results for aggregate state variables are documented in Table A6-A9. Market fixed effects are included in the autoregressions because of the considerable size difference across markets: the largest five markets have three times as many listings as the smallest five markets. As shown in the first two columns of Table A6-A9, the autoregressive coefficient exhibits significant bias without market fixed effects.

We estimated six different specifications for each aggregate state variable. The first four columns are estimated by OLS. Column (1) is a simple AR(1) regression, column (2) adds market fixed effects, column (3) and (4) introduce a separate constant before and after 2005. Column (5) and (6) are estimated by the Arellano-Bond GMM estimator that differences out market fixed effects and uses the levels and differences of the dependent variable in previous periods as instruments. Column (6) is similar to column (5), except that it also contains trend-break dummies. Omission of fixed effects tends to over-estimate the autoregressive coefficient, while OLS with fixed effects tends to generate a downward bias. The pattern that the Arellano-Bond estimate is in between OLS estimates with and without fixed effects holds across all but one specification in Table A6-A9. Our preferred specification is column (6).

We experimented with different autoregressive coefficients before and after 2005 in addition to the trend-break dummies. There is little improvement in R^2, and the slope coefficients post 2005 are unstable.

2.3 Estimates of β

We use a data dependant approach to determine the number of spline basis functions that approximate the value function. Specifically, we estimate the dynamic model several times with an increasing number of spline functions until parameter estimates stabilize, where the element by element difference between two sets of estimates β^k and β^{k-1} is smaller than half of the standard deviation:

$$k = \min\left\{ \tilde{k} \ : \ |\hat{\beta}_j^{\tilde{k}} - \hat{\beta}_j^{\tilde{k}-1}| \leq 0.5 * \text{std}\left(\hat{\beta}_j^{\tilde{k}}\right) \ , \ \forall j \right\}.$$

As described in BP1, we use MARS and the revenue function to obtain a set of splines, and use non-parametric bootstraps to estimate parameter standard errors.

For a given tolerance level, say 10^{-4}, MARS generates a set of splines ranked in decreasing importance in fitting the revenue function. We start from 24 splines and add the number of basis functions at an increment of 3: $\{24, 27, 30, ...\}$. There are no noticeable differences in parameters with 39 or more splines. These estimates, together with their standard errors, are reported in Table A10.

3 Counterfactual Analyses

The core component of the counterfactual analysis in BP1 involves solving for the fixed point of the following equation:

$$E\left(L'\right) = \sum_i \Pr(\text{active}_i; L', B') \exp(\overline{X_i^L \theta^L}). \tag{2}$$

where the summation is over incumbents and potential entrants (we assume there are $\bar{N}^E$ entrants in each period). In our counterfactuals, all 'shocks' (e.g., reduction in the commission rate) are introduced in the beginning of the sample period (year 1998), with entries and exits in period t carrying over to all future periods. Expanding equation (2) for each year in our sample, we have:

$$\begin{aligned} E(L'_{99}(S)) &= \sum_i \Pr(\text{active}_{i,99}) \exp(\overline{X_{i,99}^L \theta^L}), \\ E(L'_t(S)) &= \sum_i \Pr(\text{active}_{i,99}) ... \Pr(\text{active}_{i,t}) \exp(\overline{X_{i,t}^L \theta^L}), \quad \text{for } t > 99. \end{aligned} \tag{3}$$

which makes explicit the dependence of future periods' competition intensity L'_t on past entry and exit.

Ideally, one should sum the expected skill measure $Pr(\text{active}_{i,99})...\Pr(\text{active}_{i,t})\exp(\overline{X^L_{i,t}\theta^L})$ over all agents who were active at the *beginning* of the sample together with all potential entrants. However, agents who exit at period t are no longer observed afterward. In addition, we could not directly use information on observed entrants because the set of entrants in the data is likely to differ from that in the counterfactual. To deal with these two issues, we distribute the skill measure of exiting agents to remaining ones, and assigns appropriate weights to entrants.[1] Consider a simple two-period example with $\{A, B, C\}$ in 1999 and $\{A, B, D, E\}$ in 2000. In this example, C exits in 2000 and D and E enter in 2000. We would like to compute L'_{99} and L'_{00}.

We first obtain conditional choice probabilities $\Pr(\text{stay}_{i,t})$ and $\Pr(\text{enter}_t)$ for all agents in both periods as explained in BP1 (using a new payoff function and value function consistent with the counterfactual). Then we compute L'_{99} using equation (3):

$$\begin{aligned}
L'_{99} &= \sum_{i\in\{A,B,C\}} \Pr(\text{stay}_{i,99})\exp(\overline{X^L_{i,99}\theta^L}) + \bar{N}^E \Pr(\text{enter}_{99})\exp(\overline{X^L_{99}\theta^L}) \\
&= \sum_{i\in\{A,B\}} \left\{ \Pr(\text{stay}_{i,99}) + \frac{\Pr(\text{stay}_{C,99})\exp(\overline{X^L_{C,99}\theta^L})}{2\exp(\overline{X^L_{i,99}\theta^L})} \right\} \exp(\overline{X^L_{i,99}\theta^L}) \\
&\quad + \bar{N}^E \Pr(\text{enter}_{99})\exp(\overline{X^L_{99}\theta^L}).
\end{aligned}$$

Note that in the second equation, we distribute agent C's expected skill measure, $\Pr(\text{stay}_{C,99})\exp(\overline{X^L_{C,99}\theta^L})$, evenly to A and B. The first term in the second equation $\Pr(\text{stay}_{i,99}) + \frac{\Pr(\text{stay}_{C,99})\exp(\overline{X^L_{C,99}\theta^L})}{2\exp(\overline{X^L_{i,99}\theta^L})}$ is the weight for Agent A and B that allows us to compensate for the absence of C in 2000, as shown next.

The calculation of L'_{00} is slightly more complicated. First, we compute A and B's probability of staying active through 2000, incorporating the weights discussed above:

$$\left\{ \Pr(\text{stay}_{i,99}) + \frac{\Pr(\text{stay}_{C,99})\exp(\overline{X^L_{C,99}\theta^L})}{2\exp(\overline{X^L_{i,99}\theta^L})} \right\} * \Pr(\text{stay}_{i,00}), \text{ for } i = A, B.$$

Second, to address the issue that the number of observed entrants (two in this example) is likely to differ from the model's prediction for the counterfactual (which is $\bar{N}^E \Pr(\text{entry}_{99})$), we distribute $\bar{N}^E \Pr(\text{entry}_{99})$ equally to entrants D and E so that the total weight they carry is consistent with

[1] Another solution is to fill in missing values for all exiting agents and potential entrants by simulating s_{it}'s future path using its transition process and its observed distribution in the initial period. This would create a balanced panel for all agents. One criticism of this approach is that agents who exit are likely to have a different AR(1) process from remaining incumbents. In addition, there is a large number of entries and exits in our sample. Generating a balanced panel requires filling missing values for 60% of the sample. We do not pursue this approach given these concerns.

the model's prediction. With all agents re-weighted, we calculate L'_{00} via the following:

$$\begin{aligned} L'_{00} &= \sum_{i=A,B} \left\{ \Pr(\text{stay}_{i,99}) + \frac{\Pr(\text{stay}_{C,99}) \exp(\overline{X^L_{C,99}\theta^L})}{2\exp(\overline{X^L_{i,99}\theta^L})} \right\} \Pr(\text{stay}_{i,00}) \exp(\overline{X^L_{i,00}\theta^L}) + \\ &\quad \sum_{i=D,E} \frac{\bar{N}^E \Pr(\text{entry}_{99})}{2} \Pr(\text{stay}_{i,00}) \exp(\overline{X^L_{i,00}\theta^L}) + \bar{N}^E \Pr(\text{entry}_{00}) \exp(\overline{X^L_{00}\theta^L}). \end{aligned}$$

The calculation of L' in examples with more agents and a longer period is similar.

4 Evaluating Value Function Approximation

In this section, we briefly describe the value function approximation, compare our approach with two other commonly used methods in the dynamic estimation literature, and evaluate parameter bias associated with the value function approximation using Monte-Carlo simulations.

4.1 Value function approximation

Recall that we approximate the value function $V(S)$ by a series of basis functions $u_j(S)$:

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

and choose $\{b_j\}_{j=1}^J$ to best-fit this non-linear equation in "least-squared-residuals":

$$\left\{\hat{b}_j\right\}_{j=1}^{J} = \arg\min_{\{b_j\}} \left\| \sum_{j=1}^{J} b_j u_j(S_{(n)}) - \log\left(1 + \exp\left[\bar{R}(S_{(n)},\beta) + \delta \sum_{j=1}^{J} b_j Eu_j(S'|S_{(n)})\right]\right) \right\|_2, \quad (4)$$

where $\left\{S_{(n)}\right\}_{n=1}^N$ denotes state values observed in the data, and $\|\cdot\|_2$ is the L^2 norm.

Calculating EV involves a high-dimensional integral when there are a large number of state variables: $EV(S'|S) \simeq \sum_{j=1}^J b_j * Eu_j(S'|S)$. Note that since $u_j(S)$ is known, we can pre-compute $Eu_j(S'|S)$. This is a useful property, because recomputing a high-dimensional integral in each iteration of the parameter estimation is time-intensive. There are many existing methods for numerical integration. We choose a Quasi-Monte Carlo method because it is easy to implement and we can use the number of simulations to directly control the variance of the numerical integration. 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 Quasi-Monte Carlo method uses carefully selected deterministic sequences of points to increase the integration accuracy: the approximation error using N points is $O(\frac{1}{N})$, rather than $O\left(\frac{1}{\sqrt{N}}\right)$ as the standard Monte-Carlo numeric integration (for details, see Bretz and Genz (2009).) We use 20,000 Richtmyer points.

4.2 Value function comparison

All of the Monte-Carlo simulations discussed in this section are designed to closely mimic the empirical analysis in BP1, with the parameters in the revenue function and state transition process similar to the estimates reported in BP1. All state variables are standardized.

We compare the value function approximated by basis functions $V^B(S) \simeq \sum_{j=1}^{J} b_j u_j(S)$ with two other value functions that are popular in the literature. The first is calculated using fixed point iteration V^F (Rust 1987), while the second is calculated via Pakes & McGuire stochastic algorithm V^S (Pakes and McGuire 2001). We have conducted extensive analysis using one to four state variables. We are constrained to four state variables, because we ran out of memory on a server with 32GB of RAM when we experimented with five state variables each with 10 grid points. We focus on HP, Inv, L, and s_i (skill of agent i at time t), and fix state variable B at its sample mean and the dummy variable $L05$ at 0. All Monte-Carlo simulations discussed below use four state variables, as the discrepancy between V^B and V^F/V^S increases with the number of state variables and is the largest at four.

Both V^F and V^S require discretizing the state space. We use 10 grids for each of the four state variables. To increase the accuracy of V^F/V^S, we adjust the end points to cover 97.5% of the (random) state variable values, so that the grid is finer for state variables with less variation. Specifically, the end points for HP, Inv, L,and s_i is $[-0.18, 1.49]$, $[-1.63, 1.07]$, $[-0.51, 0.80]$, and $[-1.15, 1.11]$, respectively. We divide them into 9 evenly-spaced intervals and use the mid-points of these intervals (as well as $\pm\infty$) to form grids. For example, the grids for HP is: $\{(-\inf, -0.09], (-0.09, 0.10], ..., (1.39, \inf)\}$. With four state variables, there are 10^4 grids.

Solving the value function using Rust's fixed point iteration is straightforward. We start with $V^{F,0}(S) = \frac{\bar{R}(S,\beta)}{1-\delta}$, and iterate until the maximum absolute difference between adjacent iterations is less than 10^{-6} :

$$|V^{F,k}(S) - V^{F,k-1}(S)|_\infty < 10^{-6}.$$

Using the relative absolute difference $\left|\frac{V^{F,k}(S)-V^{F,k-1}(S)}{V^{F,k}(S)}\right|_\infty < 10^{-6}$ as the convergence criterion leads to very similar results.

The Pakes and McGuire stochastic algorithm is slightly more evolved. We start with an initial draw of the state S_0, as well as an initial guess of the continuation value for all state points $\left\{w^0(S_n) = EV(S'|S_n)\right\}_{n=1}^{10^4}$, which we set to $\frac{\bar{R}(S_n,\beta)}{1-\delta}$. Then we draw a vector of normals η_1 and compute the state of the following period: $S_1 = T * S_0 + \eta_1$. To update $w(S_0)$, the continuation value function at S_0, we first evaluate the value function at *the state of the following period* S_1 :

$$V(S_1) = \log\left(1 + \exp\left[\bar{R}(S_1, \beta) + \delta w^0(S_1)\right]\right).$$

Then we update $w(S_0)$ via the following:

$$w^1(S_0) = \frac{h(S_0)}{1 + h(S_0)} w^0(S_0) + \frac{1}{1 + h(S_0)} V(S_1),$$

where $h(S_0)$ is the number of times the state S_0 is visited prior to this random simulation. Essentially, this formula implies that $w^r(S_n)$ is a simple average of the value function evaluated at states that directly follow from S_n: $w^r(S_n) = \frac{1}{h(S_n)} \sum_i V(S_i|S_n)$, where $h(S_n)$ is the total number of times that S_n is visited prior to the rth random draw. We iterate this procedure for 1 billion draws. As in Pakes and McGuire (2001), we repeat 10 shorter-runs (with 10 million draws each) before starting the long run (with one billion draws). Some of the states are rarely updated. When we compare V^S with V^F and V^B, we weight the differences by the number of times a state is visited.

There are several factors that contribute to differences between V^B, V^F, and V^S. Discretizing the state space introduces a discretization error in V^F and V^S (both $V^F(S_n)$ and $V^F(S_n)$ are averages of the value function within a high-dimensional cube that contains the grid point S_n), the stochastic algorithm introduces simulation errors in V^S, and basis functions introduce approximation errors in V^B. All of these errors increase with the dimensionality of the state space. To examine the accuracy of our basis approximation, we report differences between V^B and V^F, V^S using four state variables in Table A11.

The results are encouraging. All three value functions are very similar, with the correlations of all three pairs of value functions exceeding 0.9997. V^B is slightly bigger than the other two, and the weighted average absolute difference between (V^B, V^F) and (V^B, V^S) is 0.0405 and 0.0402, respectively. The weight is the frequency each state is visited, $h(S_n)$, obtained in constructing V^S. The unweighted absolute difference is similar, only slightly bigger.

There is a big difference in their computational cost, however. Rust's fixed point iteration is fast and takes less than a minute, but the memory requirement increases exponentially with the number of state variables. The stochastic algorithm does not require as much memory, but it takes much longer to estimate. In our application, it takes more than two hours to simulate one billion draws (not including the initial burn-in time for ten shorter runs). If we need to re-compute the value function in each iteration of the parameter estimation, it could take weeks or months to estimate the model once. In comparison, the sieve estimation is computationally fast (less than 0.1 minute with 4 state variables and 35 basis functions) and requires little memory. The only elements stored in memory are the basis functions and their coefficients.

These advantages of the approximation method come with a cost. By replacing the true value function $V(S)$ with the approximating basis functions $\sum_{j=1}^{J} b_j u_j(S)$, it introduces the approximation error. In the following section, we evaluate the magnitude of parameter bias as a result of the approximation error.

4.3 Parameter estimates in Monte Carlo simulations

To evaluate the magnitude of parameter bias, we need to generate pseudo datasets using the true value function $V(S)$. This is an unknown object. For results reported in the first panel of Table A12, we assume that the true value function is V^F, and simulate data using the following equation:

$$W_i(S) = 1\left[\bar{R}(S,\beta) + \delta EV^F(S'|S) + \varepsilon_{i1} > \varepsilon_{i0}\right], \qquad i = 1, ..., 2500,$$

where $W_i = 1$ denotes agent i is active, $\bar{R}$ is the same as in BP1, and ε_{i1} and ε_{i0} are i.i.d. extreme value random variables. We fix the sample size of each simulated data set at 2,500, which is quite small compared with the dataset in BP1 that has more than 40,000 observations. Since the empirical exercise in BP1 has five continuous state variables and one discrete state variable (i.e., the state space is larger), we deliberately choose a smaller sample size here so that the parameter bias of the empirical estimation in BP1 is likely to be of the same magnitude as that reported here.

As described in BP1, we estimate β using constrained MLE:

$$LL(S;\beta,b) = \sum_{i,t} 1\,[\text{stay=0}] * \log[1 - \Pr(\text{stay}_{it}|\beta, b)] + \sum_{i,t} 1\,[\text{stay=1}] * \log[\Pr(\text{stay}_{it}|\beta, b)], \quad (5)$$

subject to the constraint that spline coefficients $\{b_j\}_{j=1}^{J}$ minimize the Bellman violation as specified in equation (4). We estimate β three times, with an increasing number of spline terms going from 13 to 27. Standard errors are calculated using 100 Monte-Carlo simulations. The top panel of Table A12 reports the mean and standard deviation of these β estimates.

We repeat this exercise in the bottom panel of Table A12, except that the data are generated using V^B that is calculated using the same set of spline terms as that used in estimation. Hence, there are no approximation error in these β estimates, which are standard constrained MLE estimates. We use $\hat{\beta}$ in the bottom panel as a benchmark for results in the top panel.

The Monte Carlo results reported in Table A12 suggest that the approximation error in V^B leads to little parameter bias. In the top panel, the parameter bias is around 0.003 for β_2, the revenue coefficient, when we approximate the value function using 13 spline terms. It varies between 0.002 to 0.006 with different number of basis functions. The bias for β_1 is also reasonably small, around 0.01 to 0.02. In all cases, the finite sample bias is small compared to the standard deviation of these parameter estimates, which is about 0.05-0.07. The bias in the top panel also compares favorably with the finite sample bias exhibited in the bottom panel, which varies between 0.001 to 0.02 for β_1 and 0.003-0.013 for β_2.

With a finite sample, the number of spline terms k plays an important role. In theory, more spline terms lead to a more accurate approximation of the value function at the cost of higher variances. In practice, too many spline terms often lead to various numerical problems, including difficulties in minimizing high-dimensional nonlinear functions, as well as estimation problems like collinearity and a large number of nuisance parameters. We propose a data driven method to determine k. Let $\hat{\beta}^k$ denote the parameter estimates associated with k spline terms. We increase k until the difference between $\hat{\beta}^k$ and $\hat{\beta}^{k-1}$ is smaller than half of the standard deviation of $\hat{\beta}^k$ element by element (which can be estimated using non-parametric bootstrap simulations).

We have estimated many variants of our model with one to four state variables. The estimated parameters converge fairly quickly. With 2,500 observations, the parameters often settle down when the number of spline terms increases to 10-20. In general, the bias is small and in most cases smaller than parameters' standard deviations. 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.

We take these results as evidence that approximating the value function using suitable basis functions works well in our application. Finally, our experience suggests that using the revenue function to obtain a set of basis functions that capture the shape of the value function seems very important. Choosing a dense product of polynomial terms of the state variables tends to generate noticeable bias in parameter estimates.

References

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

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

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

Table A1: Summary Statistics

Variable	Obs	Mean	Std. Dev.	Min	Max
Year	41856	2003.38	2.59	1999	2007
Skill	41856	0.08	0.49	-0.48	1.04
Exp6	41856	0.57	0.50	0	1
Active	41856	0.88	0.32	0	1
Exit	41856	0.12	0.32	0	1
PrSld	33670	0.69	0.31	0	1
Male	41856	0.35	0.46	0	1
Century21	41856	0.18	0.39	0	1
Coldwell Banker	41856	0.20	0.40	0	1
ReMax	41856	0.09	0.28	0	1
Aggregate state variables					
HP	279	0.00	1.00	-1.27	3.89
Inv	279	0.00	1.00	-1.50	3.85
L	279	0.00	1.00	-1.19	3.47
B	279	0.00	1.00	-1.17	3.63
L05	279	0.67	0.47	0	1
Ge05	279	0.33	0.47	0	1

Data source: MLS.

Table A2: Listing Share Regressions

	(1)	(2)	(3)	(4)
Skill	1.27***	1.27***	1.25***	1.27***
	(0.01)	(0.01)	(0.01)	(0.01)
Male		0.03***		
		(0.01)		
Exp6			0.14***	
			(0.01)	
Century 21				0.02*
				(0.01)
Coldwell Banker				-0.03***
				(0.01)
ReMax				0.05***
				(0.01)
Preferred Specification	X			
Estimation Method	OLS	OLS	OLS	OLS
N	32237	32237	32237	32237
R^2 adjusted	0.44	0.44	0.44	0.44

Table A3: Buying Share Regressions

	(1)	(2)	(3)	(4)
Skill	0.90***	0.90***	0.91***	0.89***
	(0.01)	(0.01)	(0.01)	(0.01)
Male		-0.06***		
		(0.01)		
Exp6			-0.10***	
			(0.01)	
Century 21				0.04***
				(0.01)
Coldwell Banker				0.08***
				(0.01)
ReMax				0.11***
				(0.01)
Preferred Specification	X			
Estimation Method	OLS	OLS	OLS	OLS
N	30986	30986	30986	30986
R^2 adjusted	0.30	0.30	0.30	0.30

Note: '*' significant at 10% level, '**' significant at 5% level, and '***' significant at 1% level. 'Skill' is agent i's number of transactions in the previous year. 'Exp6' is one for agents with six or more years of experience.

Table A4: Sold Probability Regressions

	(1)	(2)	(3)	(4)	(5)	(6)
Inv	-0.52***	-0.41***	-0.41***	-0.40***	-0.41***	-0.35***
	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.01)
Skill	0.13***	0.13***	0.15***	0.10***	0.13***	0.21***
	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)
L05		1.11***	1.18***	0.92***	1.06***	1.27***
		(0.01)	(0.01)	(0.01)	(0.01)	(0.03)
Ge05		0.75***	0.83***	0.56***	0.69***	0.83***
		(0.01)	(0.01)	(0.01)	(0.01)	(0.03)
Male			-0.20***			
			(0.01)			
Exp6				0.28***		
				(0.01)		
Century 21					0.08***	
					(0.01)	
Coldwell Banker					0.20***	
					(0.01)	
ReMax					0.03***	
					(0.01)	
Constant	1.00***					
	(0.01)					
Preferred Specification						X
Estimation Method	MLE	MLE	MLE	MLE	MLE	MLE
Market Fixed Effects	No	No	No	No	No	Yes
N	32237	32237	32237	32237	32237	32237
Pseudo R^2 adjusted	0.16	0.17	0.17	0.17	0.17	0.18

Table A5: Skill Autoregressions

	(1)	(2)	(3)	(4)	(5)	(6)
lag_Skill	0.75***	0.75***	0.75***	0.75***	0.75***	0.74***
	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)
Male		0.00				
		(0.00)				
Exp6			0.02***			
			(0.00)			
Century 21				0.01		
				(0.01)		
Coldwell Banker				0.00		
				(0.01)		
ReMax				0.02***		
				(0.01)		
L05					0.04***	0.09***
					(0.00)	(0.01)
Ge05					0.00	0.05***
					(0.00)	(0.01)
Constant	0.03***	0.03***	0.01***	0.02***		
	(0.00)	(0.00)	(0.00)	(0.00)		
Preferred Specification					X	
Estimation Method	OLS	OLS	OLS	OLS	OLS	OLS
Market Fixed Effects	No	No	No	No	No	Yes
N	30648	30648	30648	30648	30648	30648
R^2 adjusted	0.54	0.54	0.54	0.54	0.59	0.59

Note: '*' significant at 10% level, '**' significant at 5% level, and '***' significant at 1% level. 'Inv' is the sales-inventory ratio; 'L05' and 'Ge05' are indicators for year<2005 and year>=2005, respectively. See Table A3 for the explanation of Skill and Exp6. 'R^2 adjusted' in Table A4 is calculated by authors.

Table A6: Market Level Housing Value Autoregressions

	(1)	(2)	(3)	(4)	(5)	(6)
lag_HP	0.96***	0.76***	1.00***	0.87***	0.84***	0.74***
	(0.02)	(0.03)	(0.02)	(0.05)	(0.02)	(0.05)
L05			0.15***	0.04		0.29***
			(0.02)	(0.09)		(0.03)
Ge05			-0.11***	-0.12		0.17***
			(0.03)	(0.09)		(0.05)
Constant	0.06***	-0.06			0.06***	
	(0.02)	(0.09)			(0.01)	
Preferred Specification						X
Estimation Method	OLS	OLS	OLS	OLS	GMM-IV	GMM-IV
Market Fixed Effects	No	Yes	No	Yes	Yes	Yes
N	279	279	279	279	279	279
R^2 adjusted	0.92	0.93	0.94	0.93	0.93	0.93

Table A7: Sales-Inventory Ratio Autoregressions

	(1)	(2)	(3)	(4)	(5)	(6)
lag_Inv	1.02***	0.76***	0.76***	0.57***	0.91***	0.65***
	(0.05)	(0.06)	(0.05)	(0.06)	(0.04)	(0.05)
lag_HP	0.09**	0.73***	0.03	0.35***	0.38***	0.21***
	(0.04)	(0.08)	(0.03)	(0.09)	(0.05)	(0.06)
L05			-0.10**	-0.44**		-0.10**
			(0.04)	(0.18)		(0.05)
Ge05			0.74***	0.35**		0.62***
			(0.06)	(0.17)		(0.07)
Constant	0.23***	0.18			0.21***	
	(0.04)	(0.19)			(0.03)	
Preferred Specification						X
Estimation Method	OLS	OLS	OLS	OLS	GMM-IV	GMM-IV
Market Fixed Effects	No	Yes	No	Yes	Yes	Yes
N	279	279	279	279	279	279
R^2 adjusted	0.67	0.72	0.76	0.78	0.70	0.77

Note: '*' significant at 10% level, '**' significant at 5% level, and '***' significant at 1% level. 'HP' is the product of the aggregate number of house listings and the average housing price index. 'Inv' is the sales-inventory ratio. 'L05' and 'Ge05' are indicators for year<2005 and year>=2005, respectively. GMM-IV refers to the Arellano-Bond estimator. 'R^2 adjusted' in columns (5)-(6) is calculated by authors.

Table A8: Listing Share Inclusive Value Autoregressions

	(1)	(2)	(3)	(4)	(5)	(6)
lag_L	0.97***	0.56***	0.97***	0.56***	0.79***	0.79***
	(0.02)	(0.03)	(0.02)	(0.03)	(0.02)	(0.02)
lag_HP	0.10***	0.45***	0.10***	0.44***	0.36***	0.35***
	(0.02)	(0.03)	(0.02)	(0.03)	(0.02)	(0.02)
lag_Inv	-0.07***	-0.05***	-0.08***	-0.09***	-0.09***	-0.13***
	(0.02)	(0.01)	(0.02)	(0.02)	(0.01)	(0.02)
L05			0.08***	-0.03		0.03*
			(0.02)	(0.07)		(0.02)
Ge05			0.09**	0.08		0.12***
			(0.04)	(0.06)		(0.03)
Constant	0.08***	0.05			0.06***	
	(0.02)	(0.06)			(0.01)	
Preferred Specification						X
Estimation Method	OLS	OLS	OLS	OLS	GMM-IV	GMM-IV
Market Fixed Effects	No	Yes	No	Yes	Yes	Yes
N	279	279	279	279	279	279
R^2 adjusted	0.93	0.97	0.93	0.97	0.96	0.96

Table A9: Buying Share Inclusive Value Autoregressions

	(1)	(2)	(3)	(4)	(5)	(6)
lag_B	0.95***	0.48***	0.95***	0.47***	0.76***	0.76***
	(0.02)	(0.03)	(0.02)	(0.03)	(0.02)	(0.02)
lag_HP	0.10***	0.53***	0.10***	0.52***	0.36***	0.35***
	(0.02)	(0.04)	(0.02)	(0.04)	(0.02)	(0.02)
lag_Inv	-0.09***	-0.07***	-0.08***	-0.09***	-0.11***	-0.13***
	(0.02)	(0.01)	(0.02)	(0.02)	(0.01)	(0.02)
L05			0.08***	-0.08		0.04**
			(0.02)	(0.07)		(0.02)
Ge05			0.04	-0.03		0.09***
			(0.04)	(0.06)		(0.03)
Constant	0.07***	-0.05			0.05***	
	(0.02)	(0.06)			(0.01)	
Preferred Specification						X
Estimation Method	OLS	OLS	OLS	OLS	GMM-IV	GMM-IV
Market Fixed Effects	No	Yes	No	Yes	Yes	Yes
N	279	279	279	279	279	279
R^2 adjusted	0.93	0.97	0.93	0.97	0.96	0.96

Note: '*' significant at 10% level, '**' significant at 5% level, and '***' significant at 1% level. 'L' is the listing share inclusive value, and 'B' is the buying share inclusive value. 'L05' and 'Ge05' are indicators for year<2005 and year>=2005, respectively. See Table A7 for the definition of 'HP' and 'Inv'. GMM-IV refers to the Arellano-Bond estimator. 'R^2 adjusted' in columns (5)-(6) is calculated by authors.

Table A10: Parameter Estimates for Eight Sets of Spline Basis Functions

	Coefficient Estimates								Standard Errors							
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)	(14)	(15)	(16)
Revenue coefficient β_1																
	2.21	2.22	2.23	2.11	2.12	2.12	2.13	2.13	0.05	0.05	0.04	0.04	0.04	0.04	0.04	0.04
Market specific constant β_2																
ARLINGTON	-0.98	-0.98	-0.96	-0.90	-0.91	-0.90	-0.77	-0.78	0.12	0.12	0.12	0.11	0.12	0.11	0.11	0.11
BROOKLINE	-1.56	-1.54	-1.45	-1.38	-1.38	-1.38	-1.38	-1.38	0.09	0.09	0.04	0.04	0.04	0.04	0.04	0.04
CAMBRIDGE	-1.66	-1.67	-1.67	-1.46	-1.46	-1.46	-1.45	-1.45	0.05	0.05	0.05	0.05	0.05	0.05	0.05	0.05
CONCORD	-1.73	-1.75	-1.78	-1.76	-1.76	-1.77	-1.78	-1.78	0.05	0.05	0.05	0.05	0.05	0.05	0.05	0.05
DANVERS	-0.66	-0.66	-0.65	-0.67	-0.66	-0.66	-0.62	-0.63	0.11	0.11	0.11	0.11	0.10	0.10	0.12	0.12
DEDHAM	-1.03	-1.02	-1.01	-0.99	-0.99	-0.98	-1.14	-1.15	0.10	0.10	0.10	0.10	0.10	0.10	0.10	0.09
HINGHAM	-1.03	-1.03	-1.20	-1.17	-1.19	-1.19	-1.21	-1.21	0.09	0.09	0.03	0.03	0.03	0.03	0.03	0.03
LEXINGTON	-1.07	-1.07	-1.34	-1.28	-1.28	-1.28	-1.28	-1.28	0.05	0.05	0.03	0.03	0.03	0.03	0.03	0.03
LYNN	-0.97	-0.96	-0.95	-0.80	-0.80	-0.80	-0.80	-0.81	0.05	0.05	0.05	0.06	0.06	0.06	0.06	0.06
MALDEN	-1.01	-0.97	-0.96	-0.85	-0.86	-0.86	-0.83	-0.83	0.07	0.05	0.05	0.05	0.05	0.05	0.05	0.05
MARBLEHEAD	-1.03	-1.03	-1.02	-0.96	-0.95	-0.94	-0.90	-0.91	0.12	0.12	0.12	0.12	0.12	0.12	0.13	0.13
MEDFORD	-1.09	-1.08	-1.07	-1.05	-1.05	-1.04	-1.04	-1.05	0.12	0.12	0.12	0.12	0.12	0.12	0.12	0.12
NEEDHAM	-0.96	-1.27	-1.28	-1.24	-1.25	-1.34	-1.36	-1.36	0.13	0.04	0.04	0.04	0.04	0.04	0.04	0.04
NEWTON	-1.31	-1.32	-1.33	-1.31	-1.31	-1.31	-1.30	-1.31	0.05	0.05	0.05	0.05	0.05	0.05	0.05	0.05
PEABODY	-0.82	-0.83	-0.82	-0.78	-0.79	-0.79	-0.82	-0.80	0.09	0.09	0.09	0.09	0.09	0.09	0.10	0.07
QUINCY	-0.87	-0.86	-0.86	-0.68	-0.67	-0.67	-0.64	-0.65	0.05	0.06	0.05	0.06	0.06	0.06	0.06	0.06
RANDOLPH	-1.01	-1.01	-1.00	-1.00	-1.01	-1.00	-1.00	-1.01	0.12	0.12	0.12	0.12	0.11	0.12	0.12	0.12
READING	-0.87	-0.88	-0.87	-0.88	-0.88	-0.87	-0.89	-0.90	0.10	0.09	0.10	0.09	0.09	0.09	0.10	0.10
REVERE	-0.89	-0.90	-0.90	-0.63	-0.65	-0.64	-0.71	-0.72	0.06	0.06	0.05	0.07	0.06	0.06	0.06	0.06
SALEM	-0.83	-0.83	-0.82	-0.78	-0.79	-0.78	-0.88	-0.83	0.10	0.09	0.10	0.09	0.09	0.09	0.09	0.06
SOMERVILLE	-1.31	-1.31	-1.30	-1.24	-1.25	-1.24	-1.41	-1.42	0.11	0.11	0.11	0.10	0.11	0.10	0.09	0.09
STOUGHTON	-0.88	-0.87	-0.86	-0.83	-0.78	-0.78	-0.81	-0.82	0.07	0.08	0.08	0.08	0.07	0.08	0.08	0.08
WAKEFIELD	-1.04	-1.04	-1.03	-0.94	-0.94	-0.93	-0.90	-0.91	0.09	0.09	0.09	0.07	0.07	0.07	0.07	0.07
WALPOLE	-0.92	-0.92	-0.90	-0.89	-0.89	-0.89	-0.89	-0.90	0.07	0.07	0.07	0.07	0.07	0.06	0.07	0.07
WALTHAM	-0.96	-0.97	-0.96	-0.92	-0.93	-0.92	-0.95	-0.96	0.10	0.10	0.10	0.10	0.10	0.10	0.10	0.10
WATERTOWN	-0.99	-0.99	-0.98	-0.93	-1.07	-1.07	-1.07	-1.07	0.11	0.11	0.11	0.11	0.03	0.03	0.03	0.03
WELLESLEY	-1.24	-1.27	-1.30	-1.83	-1.84	-1.84	-1.86	-1.86	0.07	0.07	0.06	0.04	0.04	0.04	0.04	0.04
WEYMOUTH	-0.81	-0.80	-0.78	-0.74	-0.73	-0.73	-0.74	-0.70	0.06	0.06	0.06	0.06	0.06	0.06	0.06	0.03
WILMINGTON	-0.95	-0.95	-0.93	-0.85	-0.86	-0.86	-0.90	-0.90	0.11	0.11	0.11	0.12	0.12	0.12	0.13	0.13
WINCHESTER	-1.28	-1.37	-1.38	-1.34	-1.34	-1.47	-1.47	-1.47	0.13	0.04	0.04	0.04	0.04	0.04	0.04	0.04
WOBURN	-0.93	-0.93	-0.92	-0.80	-0.81	-0.81	-0.84	-0.84	0.13	0.13	0.13	0.15	0.15	0.15	0.16	0.17
Number of Splines	24	27	30	33	36	39	42	45								

Note: columns (1)-(8) report parameter estimates, and columns (9)-(16) present corresponding standard errors. All standard errors are calculated via 100 bootstraps. The last row is the number of spline terms used in approximating the value function.

Table A11: Comparison between V^B, V^F, and V^S

	Correlation				Num. of	Computing
	VB	VF	VS	Error	Elements	Time (min)
VB	1	0.9998	0.9997	0.023	35	0.10
VF	0.9998	1	0.99999	8.17E-07	10000	0.26
VS	0.9997	0.99999	1	0.001	10000	125.20

	Ave. Abs. Difference					
	VF	VS	Mean(V)	Std(V)	Min(V)	Max(V)
VB	0.040	0.040	1.41	1.07	0.50	7.03

Note: four continuous state variables.

Table A12: Structural Parameter Estimates

		1st Set		2nd Set		3rd Set	
Simulate Data w/ VF	β0	Mean	Std	Mean	Std	Mean	Std
beta1	1	0.990	0.07	0.982	0.07	0.984	0.07
beta2	-1	-1.003	0.05	-0.994	0.05	-0.998	0.06
Num. of basis terms		13		19		27	
Simulate Data w/ VB	β0						
beta1	1	0.999	0.10	1.017	0.06	1.020	0.08
beta2	-1	-0.987	0.08	-1.003	0.05	-1.008	0.06
Num. of basis terms		13		19		27	

Note: results of all 6 sets are calculated using 100 monte-carlo simulations. 2500 observations.