Online Appendix to "Real Estate Agents' Influence on Housing Search"

Seung-Hyun Hong
Department of Economics
University of Illinois, Urbana-Champaign
hyunhong@illinois.edu

A Details on Intrinsic Values

As defined in Section 3.2, house j's intrinsic value at period t, $h_{j,t}$, is recovered from house j's previous transaction at period t' as well as the sum of each year's housing price appreciation from period $t'+1$ to period $t-1$ in local market $l(j)$ where house j is located. I use zip code for market $l(j)$, and so $P_{l(j),k}$ is essentially time-varying zip code fixed effects capturing local housing market factors over time. To estimate $P_{l(j),k}$, I use the CoreLogic data and regress $\ln(p^S_{j,t}/p^S_{j,t'})$ on year dummies for each zip code, where these dummies are equal to 1 only for all years from $t'+1$ to $t-1$. Though this regression may be somewhat nonstandard, it is intuitive and simple. An alternative approach is to compute commonly-used repeat sales price indexes, which will produce similar estimates for $P_{l(j),k}$. $P_{l(j),k}$ is then obtained from the coefficient estimates on these dummy variables that capture the yearly average appreciation in each zip code. Once $P_{l(j),k}$ is estimated from the CoreLogic data, it can be easily matched with the MLS data, based on zip code and year.

One limitation in the MLS data is that home addresses and parcel identification numbers are inaccurate or missing in many listings, in which case previous sales prices cannot be obtained. To address this issue, I also use the CoreLogic data to obtain previous sales prices for houses in the MLS data that can be matched with the same house in the CoreLogic data. However, matching these two datasets was not straightforward, because the same issue also applies to the CoreLogic data, though this issue is more serious in the MLS data. In the end, previous sales prices can be obtained for about 40% of listings in my MLS data, which seems to be problematic. Nevertheless, this approach still results in more observations than an approach using house fixed effects – a common approach to control for unobserved house characteristics.

B Derivation of the Simplified Listing Problem

Under the exponential distribution assumption on $\nu_{j,i}$ and a functional form assumption on n_j and r_j^L, this appendix shows that

$$E\left[\max_{i\in\{1,2,\ldots,n_j\}} \nu_{j,i} \middle| I_j\right] = \omega_j - \frac{(r_j^L)^2}{2m_j},$$

which is the expression used in the second assumption in Section 4.1.

Specifically, I first assume that $\nu_{j,i}$'s are assumed to be i.i.d. random variables following the exponential distribution with its rate parameter equal to 1. This implies that

$$E\left[\max_{i\in\{1,2,\ldots,n_j\}} \nu_{j,i} \middle| I_j\right] = E(\nu_j^{n_j:n_j}|I_j) = \sum_{k=1}^{n_j} \frac{1}{k},$$

where $\nu_j^{n_j:n_j}$ denotes the maximum order statistic among n_j draws, and the second equality follows from the property of the exponential distribution (see, e.g., Arnold, et al., 1992, "A First Course in Order Statistics"). The harmonic sum above can be then approximated by the log function, because $\sum_{k=1}^{n_j} \frac{1}{k} = \ln n_j + O(1)$. Though the exponential assumption seems arbitrary, it is useful because it results in a closed form solution to $E(\nu_j^{n_j:n_j})$, which is not the case for most distributions. Moreover, the resulting expression provides an intuitive approximation of a seller's belief that n_j is positively related to sales price premiums.

I next assume that the negative relationship between n_j and r_j^L – a decrease in r_j^L is likely to bring more potential offers – can be modeled parsimoniously as follows: $x_j = \exp\left(\omega_j - \frac{(r_j^L)^2}{2m_j}\right)$ and $n_j = \lfloor x_j \rfloor$, where $\lfloor\cdot\rfloor$ is the floor function, ω_j is a parameter that may vary across housing markets, m_j is a positive variable capturing the seller's belief on house j's market, and x_j is a positive real number that links r_j^L with an integer n_j. Since plugging $\lfloor x_j \rfloor$ into $\ln n_j$ generates a discontinuous objective function, I further approximate n_j by x_j, which results in the expression at the beginning.

C Robustness Check for Instrumental Variables

I use two instruments for r_j^L in (11) and (12): house j's intrinsic value and tract-level yearly average log listing price computed by excluding house j. A potential concern about the second instrument is that listing prices of houses adjacent to house j may also affect house j's sales price premium and the probability of transaction, in which case the exclusion restriction is not satisfied. To address this concern, I include various house characteristics, zip code fixed effects, and year×month fixed

Table C1: Instrumental Variable Regression Results[a]

	sales price premium, r^S				dummy for sold listing			
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
agent-own	0.001	-0.000	0.000	0.000	-0.094**	-0.092**	-0.091**	-0.091**
	(0.002)	(0.001)	(0.001)	(0.001)	(0.009)	(0.009)	(0.009)	(0.009)
r^L	0.987**	0.988**	0.984**	0.984**	0.131**	0.101**	0.100**	0.095**
	(0.009)	(0.004)	(0.004)	(0.004)	(0.028)	(0.031)	(0.031)	(0.031)
IV: 0.1-1 mile	no	yes	no	no	no	yes	no	no
IV: 1-3 mile	no	no	yes	no	no	no	yes	no
IV: 3-5 mile	no	no	no	yes	no	no	no	yes
observations	25045	21524	21884	21884	34361	29373	29816	29816
R^2	0.927	0.925	0.927	0.927	0.051	0.051	0.051	0.052

[a]The table reports the results from instrumental variable regressions of two key dependent variables. All columns use year×month fixed effects, house characteristics, and zip code fixed effects. Columns 1 and 5 are the same as columns 1-2 in Table 4 in which the instruments for r_j^L include each house's intrinsic value and tract-level yearly average log listing price, excluding house j. Columns 2-5 (similarly, columns 6-8) also use similar instruments, except that they use the yearly average log listing prices computed by excluding houses adjacent to house j, instead of excluding only house j. Columns 2-5 and 6-8 use different definitions for those adjacent to house j: "IV: 0.1-1 mile" (or "IV: 1-3 mile") means houses located more than 0.1 mile away but within 1 mile (or more than 1 mile away but within 3 miles) from house j. Robust standard errors are in parentheses and clustered at the census tract level. + denotes significance at a 10% level, * denotes significance at a 5% level, and ** denotes significance at 1% level.

effects to ensure that the second instrument is not correlated with the error terms in (11) and (12). However, this may not fully address the concern. Therefore, I also consider modified versions of the instrument by excluding nearby houses as well. Specifically, I consider three versions. The first is to use houses located within 1 mile from house j, but exclude houses within 0.1 mile from house j ("IV: 0.1-1 mile"). The other two are similar, except that I change the distance cutoffs. The second version uses houses located more than 1 mile away but within 3 miles from house j ("IV: 1-3 mile"), while the third uses houses located more than 3 mile away but within 5 miles from house j ("IV: 3-5 mile"). These modified instruments are similar to the instruments that Bayer, et al. (2007) construct by using houses located more than 3 miles away from a given house.

The results using these instruments are reported in Table C1, where I regress r_j^S (columns 1-4) and the dummy for sold listings (columns 5-8) on r_j^L and the dummy for agent-owned listings. In the table, columns 1 and 5 are the same as columns 1-2 in Table 4 using the tract-level yearly average log listing price, excluding house j. The remaining columns use the modified instruments. The table shows that the coefficient estimates from the modified instruments do not change significantly from those using the original instruments, suggesting that the potential concern is unlikely to affect the estimates. This also supports the validity of the instruments.

The main estimation in Section 6 does not use the modified instruments for two reasons. First, the estimates from using either the original instruments or the modified instruments are similar. Second, the number of observations is reduced if the modified instruments are used as shown in Table C1, because the distance calculation requires the exact longitude and latitude of each house, but not all houses in the data can be geocoded due to errors or missing information in their addresses.

D Results from the Suburban Sample

To reduce the length of the paper, the main text reports the results only from the downtown sample. This appendix presents the results from the suburban sample covering a small part of suburban areas in the MSA studied in this paper. All tables in this appendix correspond to the tables in the main text. For example, Table D.1 below is equivalent to Table 1 in the main text, except that it reports the same summary statistics for the suburban sample, instead of the downtown sample. For most tables, the results are very similar between the downtown sample and the suburban sample, even though the comparison of Table 1 and Table D.1 shows that housing characteristics and prices are different between these two housing markets. Therefore, the results from both markets provide similar findings.

Table D1: Summary Statistics[a]

	Sold listings only	All listings (sold or withdrawn)
	(1)	(2)
listing price	546727.89	564883.94
sale price	524724.20	
sold	1.00	0.75
cumulative days on market	64.32	83.19
number of bedrooms	3.14	3.18
number of baths	1.93	1.97
house age (years)	36.63	35.44
condo	0.27	0.26
agent-owned	0.07	0.07
observations	30390	40774

[a]The table reports the mean of each variable. Column 1 uses only sold listings, while column 2 uses all sample that includes both sold listings and withdrawn listings. All prices (listing and sale) are in 2010 dollar, deflated by the Consumer Price Index. Condo is the indicator variable for whether housing tenure is condo. Agent-owned is the indicator for whether the listing was owned by an agent-seller. Sold is the dummy for whether the listing was sold.

Table D2: Price and Intrinsic Value: Agent- vs. Client-owned[a]

	Agent-owned	Client-owned
	(1)	(2)
	A. Sold listings only	
p^S (sale price)	594998.05	531334.84
intrinsic value	553421.07	532763.37
r^S (sale price premium)	1.095	1.012
observations	873	14714
	B. All listings (sold or withdrawn)	
p^L (listing price)	628542.83	561971.94
intrinsic value	559828.07	535984.88
r^L (listing price premium)	1.140	1.061
observations	1261	18729

[a]The table reports the mean of each variable among the sample for which intrinsic values can be computed. Panel A uses only sold listings. Panel B uses all sample that includes both sold listings and withdrawn listings. All prices and intrinsic values are in 2010 dollar, deflated by the Consumer Price Index. Listing price premium, r^L, is equal to listing price/intrinsic value, and sale price premium, r^S, is equal to sale price/intrinsic value.

Table D3: Regression Results[a]

dependent variable	$\ln(p^S)$	$\ln(p^S)$	r^S	r^S	r^S	$\ln(p^L)$	r^L
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
agent-owned	0.0436**	0.0089**	0.0738**	0.0758**	0.0025+	0.0332**	0.0702**
	(0.0118)	(0.0032)	(0.0080)	(0.0082)	(0.0013)	(0.0108)	(0.0091)
ln(listing price)		0.9593**					
		(0.0163)					
r^L					0.9493**		
					(0.0022)		
year×month fixed effects	yes	yes	yes	yes	yes	yes	yes
house fixed effects	yes	yes	no	no	no	yes	no
house characteristics	no	no	yes	yes	yes	no	yes
census track fixed effects	no	no	yes	no	no	no	no
zip code fixed effects	no	no	no	yes	yes	no	yes
observations	6284	6284	15587	15587	15587	10472	19990
R^2	0.989	0.999	0.102	0.089	0.972	0.988	0.093

[a]The table reports the key coefficient estimates from regressions of different price variables. The dependent variable is the log of sales price, $\ln(p^S)$, in columns 1-2; the sales price premium, r^S, in columns 3-5; the log of listing price, $\ln(p^L)$, in column 6; and the listing price premium, r^L, in column 7. House characteristics include #bedrooms, #rooms, #bathrooms, #garages, and various dummy variables for property types, basement types, and house ages. Robust standard errors are in parentheses and clustered at the census tract level. + denotes significance at a 10% level, * denotes significance at a 5% level, and ** denotes significance at 1% level.

Table D4: Instrumental Variable Regression Results[a]

dependent variable:	sales price premium, r^S	dummy for sold listing
	(1)	(2)
agent-owned	0.0018	-0.0440**
	(0.0012)	(0.0118)
r^L	0.9582**	-0.2765**
	(0.0028)	(0.0287)
year×month fixed effects	yes	yes
house characteristics	yes	yes
zip code fixed effects	yes	yes
1st stage F-stat for instruments	230.23	292.37
Sargan's J-test (p-value)	0.338	0.241
observations	15587	19990
R^2	0.972	0.097

[a]The table reports the results from instrumental variable regressions of two key dependent variables. The instruments for r^L include each house's intrinsic value and tract-level yearly average log listing price computed, excluding the house. Robust standard errors are in parentheses and clustered at the census tract level. + denotes significance at a 10% level, * denotes significance at a 5% level, and ** denotes significance at 1% level.

Table D5: Structural Estimation Results[a]

	(1)	(2)	(3)	(4)
	A. $\ln r^L$			
client-owned	-0.0610**	-0.0610**	-0.0430**	-0.0429**
	(0.0075)	(0.0075)	(0.0068)	(0.0068)
	B. $\ln r^S$			
client-owned	-0.0690**	-0.0020+	-0.0020+	-0.0007
	(0.0067)	(0.0011)	(0.0011)	(0.0012)
$\ln r^L$		0.9843**	0.9843**	0.9885**
		(0.0022)	(0.0022)	(0.0067)
	C. sold dummy			
client-owned	0.2073**	0.1668**	0.1668**	0.1405**
	(0.0399)	(0.0423)	(0.0423)	(0.0475)
$\ln r^L$		-0.7165**	-0.7165**	-1.1358**
		(0.0713)	(0.0713)	(0.1880)
	D. σ_1 for $\ln r^L$			
client-owned	-0.0228**	-0.0228**	-0.0128**	-0.0149**
	(0.0043)	(0.0043)	(0.0034)	(0.0054)
constant	0.1994**	0.1994**	0.1682**	0.1700**
	(0.0062)	(0.0062)	(0.0050)	(0.0058)
	E. σ_2 for $\ln r^S$			
client-owned	-0.0242**	-0.0033	-0.0033	-0.0139*
	(0.0065)	(0.0023)	(0.0023)	(0.0058)
constant	0.1965**	0.0347**	0.0347**	0.1694**
	(0.0071)	(0.0021)	(0.0021)	(0.0059)
	F. correlation coefficients: ρ			
$\rho_{1,2}$				0.9788**
				(0.0008)
$\rho_{1,3}$				-0.0549**
				(0.0090)
$\rho_{2,3}$				-0.0041**
				(0.0010)
year×month FE in equations A-C	yes	yes	yes	yes
house characteristics in equations A-C	yes	yes	yes	yes
zip code FE in equations A-C	yes	yes	yes	yes
instruments for $\ln r^L$ in equations B and C	no	no	yes	yes
observations	19990	19990	19990	19990

[a]The table reports the key coefficient estimates from the structural estimation, using the same observations in column 7 of Table D3. Column 1 excludes $\ln r^L$ in the equations for $\ln r^S$ (Panel B) and s_j (Panel C), and both columns 1-2 do not use instruments for $\ln r^L$. Columns 3-4 include the same instruments used in Table C1. Only column 4 allows the correlation between the error terms of the three endogenous variables. Robust standard errors are in parentheses and clustered at the census tract level. + denotes significance at a 10% level, * denotes significance at a 5% level, and ** denotes significance at 1% level.

Table D6: Model Fit[a]

	Predicted				Observed
	(1)	(2)	(3)	(4)	(5)
	A. $\exp\left[E(\ln r^S)\right]$				
Client-owned and Sold listings only	1.0559	1.0062	1.0062	0.9982	0.9956
Agent-owned and Sold listings only	1.1716	1.0885	1.0885	1.0688	1.0715
	B. $\exp\left[E(\ln r^L)\right]$				
Client-owned and Sold listings only	1.0406	1.0406	1.0374	1.0344	1.0317
Agent-owned and Sold listings only	1.1112	1.1112	1.1098	1.1051	1.1079
Client-owned and Withdrawn listings only	1.0495	1.0495	1.0611	1.0725	1.0831
Agent-owned and Withdrawn listings only	1.1255	1.1255	1.1288	1.1397	1.1322

[a]The table reports $\exp\left[E(\ln r_j^L|d_j, s_j)\right]$ and $\exp\left[E(\ln r_j^S|d_j, s_j)\right]$, where d_j is the dummy for client-owned listings, and s_j is the dummy for sold listings. Columns 1-4 report the predicted values from the estimated models, where each column in this table corresponds to the same column in Table-D5. The conditional expectation of log price premiums is computed by using the property of the bivariate normal distribution and the inverse Mills ratio. For comparison, column 5 presents the conditional mean values of observed price premiums. To be comparable with those in columns 1-4, I compute the conditional mean values of log price premiums, and then report their exponential values in column 5.

Table D7: Counterfactual Expected Values[a]

	Predicted	Counterfactual $\theta = 0$	Difference in price premium b/w agent- vs. client-owned due to agents' influence in r^L
	(1)	(2)	(3)
	A. $\exp\left[E(\ln r^S)\right]$		
Client-owned and Sold	0.9982	1.0413	61.05%
	B. $\exp\left[E(\ln r^L)\right]$		
Client-owned and Sold	1.0344	1.0794	63.65%
Client-owned and Withdrawn	1.0725	1.1190	69.20%

[a]The table reports $\exp\left[E(\ln r_j^L|d_j = 1, s_j)\right]$ and $\exp\left[E(\ln r_j^S|d_j = 1, s_j)\right]$, where d_j is the dummy for client-owned listings, and s_j is the dummy for sold listings. Column 1 is the same as column 4 of Table D6. Column 2 present the results under a counterfactual scenario where agents' influence in listing prices is removed by setting $\theta = 0$ in (10). Column 3 reports the difference between the counterfactual price premium in column 2 and the predicted price premium in column 1, relative to the difference in price premiums between agent-owned, vs. client-owned listings in column 4 of Table D6.