Data Shop

Data Shop, a department of Cityscape, *presents short articles or notes on the uses of data in housing and urban research. Through this department, PD&R introduces readers to new and overlooked data sources and to improved techniques in using well-known data. The emphasis is on sources and methods that analysts can use in their own work. Researchers often run into knotty data problems involving data interpretation or manipulation that must be solved before a project can proceed, but they seldom get to focus in detail on the solutions to such problems. If you have an idea for an applied, data-centric note of no more than 3,000 words, please send a one-paragraph abstract to david.a.vandenbroucke@hud.gov for consideration.*

Using the Health and Retirement Study To Analyze Housing Decisions, Housing Values, and Housing Prices

Hugo Benítez-Silva
The State University of New York-Stony Brook

Selçuk Eren
Levy Economics Institute of Bard College

Frank Heiland
The City University of New York-Baruch College

Sergi Jiménez-Martín
Universitat Pompeu Fabra, and FEDEA

Abstract

Few existing surveys provide detailed longitudinal information on households and their homes. This article introduces a data source, the Health and Retirement Study (HRS), which has this detailed information but has received little attention by housing researchers to date. The HRS is a rich longitudinal data set that provides information on house values, house prices, and detailed personal characteristics of those who own and sell their homes. The HRS is a nationally representative longitudinal survey that originally sampled 7,700 households headed by an individual aged 51 to 61 in the

Abstract (continued)

first interviews in 1992 and 1993. It now also samples additional cohorts of older Americans. Although the HRS is the data set of choice when analyzing the retirement behavior, savings, and health status of older Americans, given its wealth of demographic, health, and socioeconomic data, it has been rarely used to answer questions regarding the housing market. A seldom used section of the questionnaire provides detailed information about real estate transactions by households, however, enabling researchers to repeatedly observe both self-reported house values and the actual selling prices of properties sold since 1992 (originally bought in the past five decades). The article describes a number of important housing-related measures available in the HRS and illustrates the usefulness of these data by conducting a statistical analysis of the accuracy of self-reported home values. Specifically, we analyze the predictive power of self-reported housing wealth when estimating housing prices using the HRS data. The evidence shows a slight overestimation of housing values by older Americans.

Introduction

Although the Health and Retirement Study (HRS) is the longitudinal data set of choice to analyze the retirement behavior, the decisionmaking regarding Social Security as well as the savings and health status of older Americans, given its wealth of demographic, health, and socioeconomic data, it has been rarely used to analyze questions regarding the housing market.[1] A seldom-used section of the HRS, however, provides very detailed information about real estate transactions by households, which enables researchers to repeatedly observe self-reported house values, the selling prices of properties sold in the 1994-to-2008 period, and the prices originally paid as far back as the 1950s.

The HRS is a nationally representative longitudinal survey of 7,700 households headed by individuals aged 51 to 61 as of the first interviews conducted in 1992 and 1993. It has since been expanded to include even older households that were previously surveyed in the Assets and Health Dynamics Among the Oldest Old (AHEAD) and younger cohorts, such as the Children of the Depression Age (CODA) and War Babies, which refresh and complement the original sample.

This article addresses the advantages and disadvantages of using this source of data to analyze housing-related behaviors and housing market outcomes. It provides information about the instruments available in the HRS and how to construct important additional variables, questions answered by the respondents, and empirical strategies intended to overcome some of the problems with these data. The article illustrates the use of these data by presenting an interesting empirical application that analyzes the accuracy of self-reported home values and shows a slight overestimation of housing values by older Americans.

[1] See Juster and Suzman (1995) and Gustman, Mitchell, and Steinmeier (1995) for an overview of the HRS. Also see the online publication, "Growing Older in America" at http://hrsonline/isr.umich.edu.

An Unfamiliar Source of Housing Data

Using data that track particular properties over time has many advantages. For example, they enable the researcher to account for the characteristics of the houses in a detailed analysis of the dynamics of prices by regions of the country. Such surveys, however, rarely provide access to detailed information about the characteristics of the owners of those houses, their behaviors, and how much they think their houses are worth.[2] Although, at first glance, self-reported house values might not seem a key variable of interest for housing economists, it does provide essential information to researchers in a variety of fields who use household-level data and who need reliable measures of household wealth. Housing wealth is one of the pillars of the well-being of American families. It represents more than 60 percent of the average net wealth of U.S. households, according to the Federal Reserve's 2004 Survey of Consumer Finances.[3] Hence, many important decisions that households make are expected to be influenced by what they believe their houses are worth. Consequently, what the owner thinks the property is worth is very valuable information for researchers who seek to understand household decisionmaking.

What Americans think their houses are worth should be a primary concern to all housing economists, because, without understanding how the valuation evolves and how it is determined, we cannot understand the homeowners' selling decisions, both in terms of whether they decide to sell and at what price they agree to do so. Self-reported housing values may provide only a very noisy picture of the actual value of the property. It would be ideal to also have access to selling prices and compare the two measures to analyze whether reported values can be taken at face value and can be readily combined with other measures of wealth when studying the decisionmaking at the household level.

The HRS is a high-quality longitudinal data set, largely unknown in the housing literature, which provides two types of variables: (1) what individuals think their house is worth and (2) the price at which they sell the home (if a sale occurs). Up until the recent work by Benítez-Silva et al. (2009), researchers had not fully exploited the level of detail on housing wealth information available in the HRS.[4] Selected earlier research (for example, Farnham and Sevak, 2007) has used the self-reported home value information in the frequently used housing wealth section of the study but did not explore the rich data on housing transactions from the responses to the questions in the capital gains section of the HRS.

[2] Most studies use the American Housing Survey, which follows houses rather than households, or the Survey of Consumer Finances (SCF), which is not a panel data survey. See Agarwal (2007), Goodman and Ittner (1992), and Kiel and Zabel (1999). The first study on this issue was published by Kish and Lansing (1954), using the 1950 SCF, and it was not until Kain and Quigley (1972) that the assessment of self-reported home values was revisited. Kain and Quigley (1972: 803) acknowledge that "…the only accurate estimate of the value of a house is its sale price…"; however, due to data limitations and what they perceived as possibly serious selection problems, their analysis focused, as did the early study, on comparisons of households' self-reports with appraisals by experts. The latter can be considered indirect market assessments, because they use information on similar properties and try to account in the econometric study for the observable characteristics of the property.

[3] See Bucks, Kennickell, and Moore (2006). This fraction is considerably lower than in some European countries. For example, in Spain, housing wealth represents 87.5 percent of net wealth.

[4] Venti and Wise (2002) and Farnham and Sevak (2007) analyze the role of housing in retirement decisions, using the information in the housing wealth section.

This article takes advantage of the detailed information on housing transactions and valuation provided in the HRS by combining data from the wealth section with information from the largely unfamiliar capital gains section. It is important to note from the outset a number of weaknesses and limitations of these data. First, the HRS was created to analyze the socioeconomic situation and decisions of older American households; therefore, it represents only that age group and that cohort, and, although it has incorporated some other cohorts, it continues to only represent those considered to be older Americans. Ideally, we would have this richness of information for a wider cross-section of the population, but the data sets designed to represent all American households, such as the Panel Study of Income Dynamics, do not have the level of detail that we are interested in here. Another weakness of the HRS, which is common to any household-level survey, is that all the information we discuss here is self-reported and, therefore, subject to measurement errors, misreporting, and possible biases. Ongoing debate continues in the literature about the usefulness and quality of self-reported data, but it is generally believed that when considerable consistency among several sections of the survey can be demonstrated, the more serious concerns about self-reporting are unlikely to dominate over the usefulness of the data.

The Information Available in the Health and Retirement Study

The housing wealth section of the HRS asks respondents about the value of their homes (and farms or mobile homes) if they were to sell them today, the mortgages on their homes (first or second mortgages), and any home equity loans, home equity lines of credit, or other debts backed by their properties. The questionnaire also asks about the price at which the home was originally purchased, the month and year of that purchase, and real estate taxes paid on the property. The key element of self-reported housing wealth information is that it directly asks heads of household to estimate their home's current selling price. We have no way to know whether the person is thinking of selling the house, or whether it is even for sale at the time of the interview. We also have no information about the quality of the individual's assessment, at least not in this section of the questionnaire. New respondents, or those who say that they moved between waves, get to answer this part again.

Researchers have typically not gone beyond these questions in the HRS, but the survey also provides detailed information about real estate transactions that happen between waves. These additional questions, however, are asked in a completely different section—the capital gains section (called the asset change section in later waves). This section not only asks about transactions on primary and secondary residences but also about the sales of business properties, other real estate, and even financial assets.

The information on housing transactions is very detailed. The survey gathers from the respondents whether the household has bought, sold, or bought and sold a property since the previous interview and, if one of the options is the case, the price at which the house was sold as well as bracketed ranges of sale prices for those who do not answer the direct amount question. Furthermore, the questionnaire also asks about the original purchase price of the home and the date of that purchase. One shortcoming of the wording of these questions is that they are not asked sepa-

rately for both primary and secondary residences, so unless we match the information against answers in the housing section, we do not know whether the transaction was made on the primary or secondary residence. This matching is possible, but it only works if the individual purchased the primary and secondary residences in different years. If the transaction happened in the same year, there is no way to definitively know which property the person is talking about without looking at the reported values and guessing for which property the person has supplied the information. The latter option is time consuming and error prone because it requires going over every questioned transaction.

The survey also asks respondents about improvements made to the properties (again not differentiating between primary and secondary residence), both in terms of whether any improvements were made and the value of that work (which includes the value of the work they might have done themselves). Overall, this survey presents a fairly detailed picture of the housing assets these older American households have and the transactions they have completed during the 1992-to-2008 period.

With all these pieces of information, we can construct a number of useful variables beyond self-reported housing values, sale prices, and original purchase prices, such as the average capital gain that households experienced on their properties, the average equity in the properties, the number of years households own a house before selling it, and information about the property owners regarding their age, marital status, race, income, and so on.

Exhibit 1 summarizes some of the characteristics of financially knowledgeable homeowners and their assets. The columns break down the sample according to the selection criteria: whether or not individuals sell their house during the 1992-to-2002 period for which we are analyzing data. Note that, given the longitudinal nature of the sample, homeowners may be observed up to six times but are asked whether they sold a house they owned at only five of those occasions.

From the 1,086 observations we have in the first six waves of the HRS that report valid positive selling prices on households' homes and, at the same time, reported a valid value of a home they previously owned, we eliminated 210 observations because we did not have valid information about when they bought that home or when they sold it. Not having information on the first variable (when they bought the home) does not allow us to match the property exactly, and not having information on the second variable (when they sold the home) prevents us from using the difference in months between the time of the self-reported value and the time they sold the property, which is an important variable in our econometric application. We also eliminated homeowners who reported a sale price 0.2 times the self-reported house value and less, or 5 times the self-reported house value and more (a total of 40 individuals). These extreme values occur mostly due to coding errors.[5]

[5] Because of all these restrictions, our estimated sample is reduced to the 836 observations used in the ordinary least squares estimations. The selection-corrected estimations use only 665 observations because we lose some observations by including home equity in the selection equation as an exclusion restriction that allows us to nonparametrically identify the selection-corrected specification.

As shown in exhibit 1, those who did not sell a house during the observed period reported lower home values, purchase prices, and capital gains. The average home tenure for sellers is shorter than for nonsellers, but it is still almost 18 years. On the other hand, nonsellers have less home equity, are less likely to be White, have lower educational attainments, and lower earnings. The marital status, average age, and gender composition are similar for both sellers and nonsellers. Looking at the sellers, we observed that self-reported home values are greater than selling prices by around 2 percent.

Exhibit 1

Summary Statistics

Variable Name	Sellers		Nonsellers	
	Mean	Standard Deviation	Mean	Standard Deviation
Selling price	140,022	114,673		
Self-reported house value	143,199	108,510	122,947	111,984
Original purchase price	79,929	85,219	56,838	74,982
Capital gains	63,269	75,570	66,109	84,833
House tenure	17.41	11.30	21.28	11.41
Home equity	103,911	98,623	96,101	95,982
Bachelor's degree	0.3779	0.485	0.28	0.448
Professional degree	0.1411	0.348	0.109	0.311
Married	0.726	0.446	0.747	0.434
White	0.886	0.317	0.782	0.412
Age	61.52	5.84	61.53	5.65
Male	0.559	0.496	0.543	0.498
Earnings	87,820	113,314	75,525	119,157
Number of observations	836		18,144	

An Empirical Application: Testing the Accuracy of Self-Reported Housing Wealth

This interesting source of data can be used in a variety of ways to supplement, complement, and even contrast information on housing values, housing prices, and characteristics of homeowners obtained in other surveys. In this section we provide a simple empirical analysis that tests within a simple regression model the accuracy of self-reported housing wealth measures in the HRS.

In the data, we observe the market value of a property when the individual reports the transaction price of a house they have sold since the last survey wave. Therefore, the self-reported house value is obtained from the previous wave of data. Given data collection every other year only, as many as 24 months may pass between the measurement of the sale price and the self-reported house value. In the interview, individuals are asked about the current market value of their homes rather than asked to forecast the price for a future period. To correct for possible bias in the estimation of the coefficient of interest resulting from possible appreciation (depreciation) of the value of the house

during that time, we control for the number of months between the observances of these two variables.[6] The ordinary least squares (OLS) specification can then be written as follows:

$$y_i^t = \beta X_i^{t-1} + \alpha T + \varepsilon_i \quad (1)$$

where X_i^{t-1} represents the self-reported house value from the previous wave, and T represents the number of months between the time the market price refers to and the time of the self-reported home value estimation. The dependent variable is the price of the property reported by the individual, and, if the homeowners predict the market value of their house accurately, we expect to find that $E[\beta_i \mid X_i^t, \varepsilon_i] = 1$. If homeowners overestimate (underestimate) the value of their home, then the estimated slope coefficient β will be less than (more than) one.[7]

One underlying concern with the OLS specification presented is that we only estimate the relationship of interest for the sample of sellers. If sellers are very different from nonsellers on unobservable characteristics, we would not be able to generalize our results to the whole population. We follow the classic work of Heckman (1979), which reformulates this selection problem as a specification bias that has as a source the omission of a variable that represents the sample selection rule. We correct this problem by adding the inverse mills ratio, which results from estimating a selection equation, into the equation of interest. This selection equation can be the result of a probit estimation if we assume a Gaussian distribution of the error term of the binary choice model of selling a property. It is common to add an exclusion restriction to this selection equation to obtain nonparametric identification of this nonlinear model (the parametric identification is guaranteed by the nonlinearity of the model), and in our case the variable we use is the home equity on the home.

Exhibit 2 presents the results from the different specifications and estimation strategies. The OLS estimate of β, the coefficient on the self-reported house value when estimated without a constant, is 0.903. This point estimate implies an overestimation of about 10 percent in house values. If we estimate the model with a constant, the coefficient of interest goes down to 0.887, but the constant is estimated as not statistically different from zero. Both specifications explain a very large proportion of the variation in selling prices, which confirms the reliability of the model we present in this article.

[6] Notice that this discrepancy in the timing of the assessment suggests that the relationship in (1) is potentially nonlinear. We have allowed for the difference in months to enter nonlinearly (which could capture changing economic conditions in the months before the sale, which could affect the price, like movements in the interest rates), but the results have not changed. One possible alternative would be to adjust all the observed prices to the same time period. This adjustment, however, may create some unwanted measurement error because, in many cases, only a few months of difference existed between reports. The results of our preferred specification remain literally unchanged; therefore, the empirical evidence suggests that those who sell shortly after the interview do not report systematically more accurate estimates of the selling price of their properties than those who sell shortly before the following interview.

[7] There is no reason to believe that the model should contain a constant, because no minimum market value exists for the houses, and the left- and right-hand sides are measuring the same asset. In fact, we have run several empirical specifications with a constant and it comes out to be insignificant, as expected, no matter how we specify the model. In the empirical work, we present results with and without a constant in the regression.

Accounting for selection, we find the coefficient of the inverse mills ratio to be statistically insignificant, suggesting that there is no evidence that sellers differ from nonsellers in unobservable ways.[8] Although the coefficient for reported house values decreases slightly, the standard errors increase.

While given the precision of our estimates, we cannot reject the hypothesis that agents are assessing the value of their houses with accuracy; the point estimates indicate considerable overestimation of the value of the properties.

One additional concern with this model, which is explored in some detail in Benítez-Silva et al. (2009), is the endogeneity of self-reported home values due to unobserved heterogeneity grounded on local market conditions and unobserved house characteristics.

Exhibit 2

The Accuracy of Self-Reported Home Values

Dependent Variable: Sale Prices	OLS		OLS, No Constant		Selection Corrected	
	Coefficient	Standard Error	Coefficient	Standard Error	Coefficient	Standard Error
Self-reported house value	0.887	0.087	0.903	0.0601	0.894	0.092
Months between the report and the sale	468.41	351.06	741.04	407.13	527.62	389.24
Constant	7,056	13,411	—		—	
Inverse mills ratio	—		—		3,277	6,748
Adjustment R-squared	0.7067		0.882		0.8763	
Observations	836		836		665	

OLS = ordinary least squares.

Conclusions

Few existing surveys enable researchers to study households and their homes over time. The purpose of this article is to introduce one data source, the Health and Retirement Study (HRS), which has this information but has received little attention by housing researchers to date. The HRS is a longitudinal data set that provides information on self-reported house values, house prices, and detailed personal characteristics of those who own and sell their homes. The HRS is well known and frequently used among researchers in the fields of aging and retirement, but its rich section on housing, covering the prices of properties sold after 1992 and the prices of properties bought as early as the 1950s, is not well known and has rarely been used. This article discusses the potential and the limitations of the housing data collected in the HRS. We describe the housing-related instruments available in the HRS and show how to construct a number of important additional measures related to housing transactions and wealth. We illustrate the usefulness of these longitudinal data for housing research by conducting a statistical analysis of the accuracy of self-reported home

[8] In a related context but estimating a different type of home sale price equation, Ihlanfeldt and Martínez-Vázquez (1986) also found no evidence of sample selection bias when estimating an equation of sale prices.

values. This application is motivated by the fact that self-reported home values are widely used as a measure of housing wealth by researchers employing a variety of data sets and studying a number of different individual and household-level decisions.

Acknowledgments

Hugo Benítez-Silva acknowledges the financial support from Grant Number 5 P01 AG022481-04 from the National Institute on Aging on a related project. Benítez-Silva and Frank Heiland thank the Michigan Retirement Research Center for their support on a related project. Sergi Jiménez-Martín and Benítez-Silva acknowledge the financial support of the Spanish Ministry of Education through project number SEJ2005-08783-C04-01. Any remaining errors are the responsibility of the authors.

Authors

Hugo Benítez-Silva is an associate professor and Director of Graduate Studies in the Department of Economics at The State University of New York-Stony Brook.

Selcuk Eren is a research scholar at the Levy Economics Institute of Bard College.

Frank Heiland is an assistant professor and CUNY Institute for Demography Faculty Associate in the School of Public Affairs at The City University of New York-Baruch College.

Sergi Jiménez-Martín is an associate professor in the Department of Economics at the Universitat Pompeu Fabra and Fundación de Estudios de Economía Aplicada.

References

Agarwal, Sumit. 2007. "The Impact of Homeowners' Housing Wealth Misestimation on Consumption and Saving Decisions," *Real Estate Economics* 35 (2): 135–154.

Benítez-Silva, Hugo, Selcuk Eren, Frank Heiland, and Sergi Jiménez-Martín. 2009. How Well Do Individuals Predict the Selling Prices of Their Homes? Unpublished paper.

Bucks, Brian K., Arthur B. Kennickell, and Kevin B. Moore. 2006. "Recent Changes in U.S. Family Finances: Evidence from the 2001 and 2004 Survey of Consumer Finances," *Federal Reserve Bulletin* 92: A1–A38.

Farnham, Martin, and Purvi Sevak. 2007. Housing Wealth and Retirement Timing: Evidence from the HRS. Working paper 2007-172 (October). Ann Arbor, MI: University of Michigan, Michigan Retirement Research Center.

Goodman, John, Jr., and John B. Ittner. 1992. "The Accuracy of Home Owners' Estimates of House Value," *Journal of Housing Economics* 4: 339–357.

Gustman, Alan L., Olivia S. Mitchell, and Thomas L. Steinmeier. 1995. "Retirement Measures in the Health and Retirement Study," *Journal of Human Resources* 30 (Supplement): S57–S83.

Heckman, James J. 1979. "Sample Selection Bias as a Specification Error," *Econometrica* 47 (1): 153–161.

Ihlanfeldt, Keith R., and Jorge Martínez-Vázquez. 1986. "Alternative Value Estimates of Owner-Occupied Housing: Evidence on Sample Selection Bias and Systematic Errors," *Journal of Urban Economics* 20 (3): 356–369.

Juster, Thomas F., and Richard Suzman. 1995. "An Overview of the Health and Retirement Study," *The Journal of Human Resources* 30 (Supplement): S7–S56.

Kain, John F., and John M. Quigley. 1972. "Note on Owners Estimate of Housing Value," *Journal of the American Statistical Association* 67: 803–806.

Kiel, Katherine A., and Jeffrey E. Zabel. 1999. "The Accuracy of Owner-Provided House Values: The 1978–1991 American Housing Survey," *Real Estate Economics* 27 (2): 263–298.

Kish, Leslie, and John B. Lansing. 1954. "Response Errors in Estimating the Value of Homes," *Journal of the American Statistical Association* 49: 520–538.

Venti, Steven F., and David A. Wise. 2002. "Aging and Housing Equity." In *Innovations in Retirement Financing*, edited by Zvi Bodie, P. Brett Hammond, and Olivia S. Mitchell. Philadelphia, PA: University of Pennsylvania Press and the Pension Research Council: 254–281.