Self-Assessed Housing Values in the American Community Survey: An Exploratory Evaluation using Linked Real Estate Records

W. Ward Kingkade
Social, Economic and Housing Statistics Division
U.S. Census Bureau[1], 4600 Silver Hill Road, Suitland, MD 20746
W.Ward.Kingkade@census.gov

Abstract

This investigation uses a linked dataset of occupied housing records from the 2009 American Community Survey (ACS) and information on the same housing units in a database of national scope obtained from administrative records. The analysis examines the difference between the self-reported home value obtained from the ACS householder and a measure of home value derived from the administrative data for single-family homes. The statistical analysis of these data relies on the R Survey package developed by Thomas Lumley (2010), which accommodates the replicate weights used for variance estimation from the ACS.

Key Words: Record Linkage, Replicate Weights

1. Background, Data, and Methods

A house is typically the most valuable asset a household possesses, and real estate, most of which is residential, is the largest single component of U.S. national wealth (DiPasquale and Wheaton, 1996). Censuses and sample surveys have served for long as fundamental sources of information on the size and composition of the nation's housing stock. As time draws on, interest has turned increasingly towards the utilization of administrative records as supplements to, and potentially less costly substitutes for, field enumeration in U.S. Federal Statistics. This outlook extends to housing statistics.

The present paper reports on an analysis of a dataset in which housing unit records from the 2009 American Community Survey (ACS)[2] have been linked to corresponding records from a nationwide database derived from county level records for property parcels assembled by CoreLogic Inc. The CoreLogic tax roll database (CL) includes a variety of attributes of the property parcel, including property value, structural characteristics of the housing unit(s) on the parcel, and information on mortgages, sales, and ownership. The matching was accomplished by the Center for Administrative

[1] This report is released to inform interested parties of ongoing research and to encourage discussion of work in progress. The views expressed on statistical, methodological, technical, or operational issues are those of the author and not necessarily those of the U.S. Census Bureau.

[2] The ACS offers two sample datasets: one for housing units, and another for individual residents. The sample of interest in the present analysis is the ACS housing unit sample.

Records Research & Applications at the U.S. Census Bureau, matching CoreLogic street addresses to the Census Bureau's Master Address File with standard software and settings. An overall match rate to 2009 ACS housing records of 80% was obtained for single-family units, which are the focus of the analysis which follows.

The purpose of the present analysis is exploratory, investigating the manner in which the addition of the CoreLogic information can lend insight into the quality and accuracy of responses obtained from ACS householders. In this analysis, the CoreLogic data are taken as a benchmark against which the responses elicited from ACS householders are evaluated[3]. This analysis compares overall home value as reported by ACS householders with the values provided for the same homes in the county records assembled by CoreLogic. The ACS value item asks the householder to indicate what he/she believes the home would sell for as of the interview date, involving a degree of subjectivity as well as potential error in recall of the actual sales price. The CL dataset provides 4 indicators of overall property value as of the most recent recording date: the assessed value as of the county's most recent tax assessment; the appraised value as of the most recent recorded appraisal; the market value, representing the state's estimate of what the property would sell for at a given date, and a "calculated value" that indicates CoreLogic's "best guess" synthesis of the former three values[4]. In many instances one or more of these values are missing from the county records or otherwise unavailable. For the present analysis, the highest of the respective values available for the parcel serves as the measure of the parcel's value[5]. In this paper we examine the factors that are associated with the difference between the ACS householder's self-assessed value and the maximum of the CL values for the property and consider possible underlying mechanisms, with a view to illustrate the utility of combining the datasets. No attempt shall be made to estimate parameters of the US housing stock or pursue hypotheses about their determinants.

For housing units, the ACS involves a 2-stage stratified sample, necessitating the use of weights for calculation of representative means and more complex statistics, as well as variance estimation (US Census Bureau, 2009 and 2010). In keeping with current Census Bureau practice, standard errors for statistical parameters are computed using replicate weights obtained from a resampling algorithm. For this reason, the present analysis has relied extensively on the Survey package in R developed by T. Lumley (2010) and made available on the CRAN website (R Core Team, 2012), which accommodates replicate weights for complex sample designs.

[3] The CoreLogic data are undoubtedly subject to errors of their own. A more thorough analysis of errors in the CoreLogic data as well as the mismatch rates is deferred here, pending a more exhaustive address matching effort.

[4] This is not to be confused with CoreLogic's proprietary estimate of property value, derived from Automated Valuation Estimators and marketed by CoreLogic Inc.

[5] A major reason for preferring the maximum to a median or mean value is the overall tendency for householders to value their homes more highly than CoreLogic, as seen below.

The present study is restricted to single-family owned homes, which comprise the bulk of the U.S. housing stock. Such units present the most straightforward cases of "homes" for analysis[6].

2. Dynamics of the U.S. Housing Market in recent decades

Historical changes in housing prices in the U.S. since the 1990s cannot be overlooked in this analysis, since many householders may be apt to recall the sale value of the home. During the 1990s homeownership and housing prices rose sharply, and mortgages on favorable terms were relatively easy to obtain. Figure 1 plots the OFHEO[7] index of housing prices for the U.S. as a whole. As can be seen, housing prices peaked in 2006 and then abruptly declined. In the preceding housing "boom", many adjustable-rate mortgages whose rates could be reset to higher levels were obtained by numerous borrowers in the expectation of continued rises in housing prices. When housing prices declined in the context of concomitant job reductions, many such borrowers were unable to make payments, defaulting and resulting in foreclosures (Weaver, 2008). This triggered a financial crisis and affected overall U.S. GDP. Only very recently, starting perhaps in 2010, have housing prices begun to rise again.

Given this background, it should not be surprising if longer-term homeowners tended to think back to an earlier prevailing level of higher home values than the 2009 market afforded. Figure 2 plots smoothed values of 2009 ACS householders' self-assessments of home value, the maximum of the CoreLogic tax roll based values, and the home's recorded sale price adjusted to the interview date by the OFHEO housing price index against sales dates in the 25-year period up through 2009. The (kernel) smoother is an indicator of the local mean as of the given sales date (Wand and Jones, 1995). The highest values throughout are the ACS self-assessments, consistent with the observation from the American Housing Survey that residents tend to overvalue their homes relative to estimates from external sources such as CoreLogic's Automated-Valuation Model estimates (Carter, 2012). Nonetheless, there is a notable similarity in the overall movements in the ACS and CL value series in the figure[8]. The principal departure in trend between the two series is the pronounced rise in self-assessed values of homes purchased in the latter portion of the 1990s and steady decline in the same series from a date prior to 2000. Intriguingly, the OFHEO-adjusted and the CL series are in close agreement from a date around 2004 onwards. The rise exhibited in the ACS and OFHEO-adjusted series is practically absent in the CL series.

3. Duration of Residence and Duration since Sale

How long ago the current home was purchased should bear a definite relationship to the accuracy of recall. In addition, the intervening dynamics of the housing market

[6] Single family homes incorporated in condominiums are among the observations included in the analysis.

[7] This is an acronym for Office of Federal Housing Enterprise Oversight, which has since 2008 been subsumed into the Federal Housing Finance Agency.

[8] It should be kept in mind that all series in Figure 2 depict estimated 2009 sales values of homes. This includes some homes with recorded sales dates that are later than the ACS interview date.

may not be well remembered by all householders, as Figure 2 has indicated. Because only about one third of the homes in our matched sample had sales dates recorded in the CL database, we prefer to investigate this association in terms of the householder's duration of residence in the home as indicated in the ACS. As Figure 3 illustrates, the average difference between the ACS and CL values for the same home increases as residence duration rises across the following categories: one year and under (Y1); over one up to five years (Y5); over 5 to 20 years (Y20); and over 20 years (Y20+), the latter mean being considerably higher than that of its neighboring category. It should be noted that the difference in the means for durations up to 1 and from 1 to 5 years is not statistically significant[9], although our interest is in the overall tendency of means across all categories, which is significant.

4. Association with Educational Attainment

Educational attainment is a characteristic of ACS householders that one might expect to be related to the accuracy of self-assessed home value. Figure 4 shows the mean ACS-CL home value differences for four categories of educational attainment: less than high school (nohs), completed high school (hs), college graduate (col), and holders of graduate degrees (grad). The value differences increase monotonically across these categories, and all differences between consecutive categories are statistically significant. Clearly, proximity of self-assessed home value to the maximum of CL values for the home does not increase with educational attainment. Perhaps education is reflecting socioeconomic status more than knowledge of the housing market. It is also conceivable that education provides an advantage in obtaining favorable terms in home purchases.

5. Variation by Race/Ancestry

Race and ancestry (in particular, Hispanic origin) is a factor associated with pervasive social and economic differences in U.S. society, which extend to the housing market. The mean ACS-CL value differences are plotted in Figure 5 for the following four groups: Black only, White only Hispanic[10], White only non-Hispanic, and all others. All differences between categories are statistically significant. The first three categories typically follow one another in succession, and an increase in home value might be expected across these categories. That the "all other" category registers the highest ACS-CL home value difference is unexpected. This residual category is quite heterogeneous, encompassing Native Americans and Alaska Natives, Asians, Pacific Islanders, and all persons who indicated affiliation with more than one race/ancestry category.

6. Multivariate Analysis

In the analysis thus far, no attempt has been made to control for potentially confounding influences of factors such as income or age. In order to address this concern, two models have been fit with ACS-CL value difference measures as the dependent

[9] Unless otherwise indicated, our criterion for statistical significance is the p=0.05 level.

[10] Please note that this category is quite different from Hispanics of all races.

variables[11]. The independent variables include household characteristics, attributes of the householder, and a contextual neighborhood variable.

Table 1 presents the results of the first regression, in which the dependent variable is the ACS-CL home value difference. All coefficients are statistically significant at the 0.05 level or better. The householder's duration of residence continues to exhibit a positive effect, as seen in the bivariate analysis. This may reflect the memory of a more prosperous past. It may also be due to cumulative inertia or selectivity for satisfied homeowners at longer durations, those who were less satisfied having relocated.

Age of householder, the closest correlate of duration of residence (r=0.59), exhibits a substantial positive net effect in Table 1. The association of age with life cycle stage undoubtedly plays a highly important role in this result. Older householders are apt to be consumers of more valuable housing as required by the demands of family growth, at least until some rather advanced age.

Household income, the obvious immediate determinant of the household's ability to purchase housing, is positively related to the ACS-CL home value difference measure. It seems eminently plausible that wealthier people would tend to value their own property more highly than those who are not as wealthy, whatever the "objective" value of the property may be. It is also plausible that subjective assessment error may operate in relative rather than absolute terms, so that those who own more valuable homes would overvalue their property by greater absolute amounts than owners of less costly dwellings.

Education, measured in years of school completed, continues to exhibit a positive effect in the multivariate analysis. The effect of education in Table 1 does not appear to be capturing information-processing competence with regard to current housing prices as represented by the degree of agreement between the self-assessed and county-record-based home value measures.

While the present analysis is limited to owned homes, there are two tenure categories distinguished in the sample analyzed: those who own their homes free and clear, and those who own with a mortgage. The dummy variable "Own Free & Clear" in Table 1 takes a value of one for the former tenure category, while the latter category is assigned the value of zero. This variable's negative net effect may indicate that, controlling for the other variables in the equation, owners who do not have a mortgage to pay up are freer than others to sell their property and more apt to be familiar with current housing values.

Another dummy variable ("HHLDREmployment") taking a value of 1 if the householder was employed full or part time and zero otherwise emerges with a strong negative effect in Table 1. This variable may be viewed as a measure of involvement in the economy, and those who are more involved would be likely to know the history of local home values more accurately than persons who are less involved, contributing to a more pessimistic self-assessment of the values of their own homes.

[11] The models were fit by the SVYGLM function of the R Survey package, which takes proper account of the replicate weights. Because the generalized model has an identity link with a Gaussian error, it amounts to a weighted least squares fit.

The four ethnic categories distinguished in Figure 5 appear in the regression equation as a set of three dummy variables ("HHLDRWonlyNonHisp", "HHLDRBonly", and "HHLDROthRace/Origin"). These take on values of 1 when the householder is a member of the following three groups, respectively: 1) White only Hispanic; 2) Black only; and 3) neither of these nor White only non-Hispanic. Otherwise these dummies take on values of zero. White Only Non-Hispanics constitute the control group. The effects of these dummies depart from the results of the earlier bivariate analysis in that they are all negative. A householder's membership in any of the three dummied categories is associated with a lower ACS-CL home value differential than what obtains for White only non-Hispanics. Black only householders undervalue their homes the most, by almost 20 thousand dollars relative to their White only non-Hispanic counterparts.

Whether the language of the household is English or something other, as reflected in the dummy variable "HHLangEnglish", has a substantial positive effect in Table 1. This effect may reflect the confidence that English fluency confers on the householder in navigating the housing market.

A dummy variable reflecting the presence of the householder's own children in the household ("HHLDRChildren") is positively associated with the ACS-CL home value difference measure. The presence of children tends to promote demand for larger homes, which are typically priced higher than smaller homes, other things being equal. It may also be that parents are apt to value characteristics of their home that are less important to householders with no children. However, the comparatively small magnitude of this effect should also be kept in mind.

Table 1 contains another dummy variable, "HHLDRForeignBorn", which indicates whether the householder was born outside the 50 U.S. states. This variable exhibits a positive net effect. Perhaps some of the foreign born may be less familiar with the U.S. housing market than persons born in the U.S.. They may value, or end up paying higher prices for, their homes than would U.S. residents born in the 50 states for a similar home.

In an effort to incorporate a contextual variable representing the local social environment, the race/ancestry composition as of Census 2010 of the ACS 2009 Census tract in which the home is located[12] is taken into account in the equation by the variable "LogitTractPWonlyNH", which is the logit of the proportion of the tract's population comprised by White only non-Hispanics. This variable has an unexpected negative effect on the ACS-CL home value difference measure. This question deserves further attention.

6.1 Relative Differences

At several points above, the idea that effects of various regressors on the ACS – CL value differences may operate in relative rather than absolute terms has been suggested. Perhaps survey respondents are apt to think of housing prices in "round numbers" or orders of magnitude. As a measure of relative difference between the ACS and CoreLogic home value measures, we employ the ratio of the arithmetic difference between the two measures to the greater of the two measures for the same home. That is,

[12] Appropriate adjustments for tract boundary changes were made in order to correspond to ACS 2009 geography.

$$Reldif = \frac{ACS - CL}{\max(ACS, CL)} \quad ,$$

where ACS and CL denote the corresponding home value measures.

Our measure of relative difference varies between a maximum of 1 and a minimum of -1, which suggests the use of a logistic model to capture the relationship of this measure with other variables. Because the relative difference measure is continuous in the interval (-1,1) and its components are observed as continuous variables in the data, it can be converted into a more statistically tractable measure by applying the generalized logit transformation,

$$\lambda(y) = \ln\left(\frac{y(x) - L}{U - y(x)} \right) \quad ,$$

Where U is the upper asymptote, L is the lower asymptote, and y is the relative difference measure. When y is a logistic function of x, its generalized logit is a linear function of x.

Table 2 presents the parameter estimates from the regression of the generalized logits of the relative difference measure on the set of independent variables present in Table 1. A number of the estimated coefficients differ from those in Table 1. In particular, age exhibits a significant negative effect on the relative differences, while the effect of duration of residence retains a highly significant positive effect. Education emerges with a significant negative effect that is consistent with the idea that education confers an ability to handle information that is beneficial in navigating the housing market. The dummies for owning without a mortgage and householder employment retain their significant negative effects. English as the household's language continues to exhibit a significant positive effect. Foreign born status has a negative net effect on the relative difference measure that contrasts with its positive effect on absolute differences. This seems to indicate that as housing value rises, the overvaluation of homes by the foreign born does not rise as much.

The pattern of effects of the three race/origin dummies in Table 2 differs from that in Table 1. Black only householders appear more prone than other householders to overvalue their homes in relative terms. Among Black only householders, owners of higher valued homes are more likely than their White only non-Hispanic counterparts to overvalue their homes in relative terms. Taken together with the negative net effect of the householder's Black only race/origin status on absolute differences, it seems that more than one direct mechanism may be responsible: absolute undervaluation may rise with home value but fail to keep pace in relative terms, absolute overvaluation may be greater in relative terms for higher valued homes than those of lower value, or both phenomena may be occurring. This warrants further attention.

The effect of the presence of householder's own children on the relative difference measure differs from its significant positive effect on absolute differences in Table 1. The sign of the effect on relative differences is negative in Table 2. However, this effect is not statistically significant.

Another noteworthy difference in net effects on absolute and relative differences between self-assessed and CoreLogic home values pertains to local ethnic composition as represented by the common logit[13] of the White only non-Hispanic share of the population of the household's census tract. In Table 2 this contextual variable has a significant positive effect on the generalized logit relative difference between ACS and CoreLogic home value measures. The discrepancy between ACS and CoreLogic values decreases or becomes increasingly negative as the local proportion White non-Hispanic rises, but evidently not in proportion to increases in home value.

7. Conclusion

There is clearly a gain from integrating the data on ACS housing records with the CoreLogic dataset. There are clearcut differences between the measures of home value from the ACS and the CoreLogic database. These are related to characteristics of the householder, the household, and the local area in which the household is located.

If the preliminary exploration above has stimulated curiosity, the present analysis has been a success. Further work unquestionably remains to be done. An analysis of structural characteristics, such as types of rooms and amenities, would be one direction to pursue. Estimating a model that incorporates indirect effects of exogenous variables on housing values through intervening characteristics of the dwelling unit might elucidate the associations detected in the preliminary analysis above. Incorporation of data for additional years would facilitate such analyses.

References

Carter, G.R., "Housing Units with Negative Equity", Cityscape, 14(1), 2012, pp. 149-165.

DiPasquale, D. and W.C. Wheaton, Urban Economics and Real Estate Markets. Upper Saddle River, NJ. Prentice Hall. 1996.

Lumley, T., Complex Surveys. Hoboken, NJ, John Wiley & Sons. 2010.

R Core Team, R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. 2012. ISBN 3-900051-07-0, URL http://www.R-project.org/.

U.S. Census Bureau, ACS Design and Methodology. US Bureau of the Census, Washington, DC. 2010.

-----, American Community Survey: Design and Methodology. Washington, DC, US Government Printing Office. 2009.

Wand, M.P. and M.C. Jones, Kernel Smoothing. Boca Raton, FL. CRC Press. 1995.

Weaver, K, "The sub-prime mortgage crisis: a synopsis". Global Securitization and Structured Finance. Deutsche Bank. 2008.

[13] The term "common logit" here refers to the logit of a proportion, which amounts to a generalized logit with asymptotes of 0 and 1.

Table 1: Regression of difference between ACS householder's self-assessed home value and maximum of available CoreLogic county record value items on selected characteristics of the householder, household, and locational context (dollars)

Variable	Coefficient	Standard Error	t	Pr(>|t|)	significance
(Intercept)	-40409.442	3784.284	-10.678	0.00000	***
HHIncome	419.837	14.129	29.715	0.00000	***
HHLDRAge	648.589	57.423	11.295	0.00000	***
Residence Duration	760.091	48.674	15.616	0.00000	***
Years of Schooling	2644.735	178.066	14.853	0.00000	***
Own Free & Clear	-13418.708	1296.424	-10.351	0.00000	***
HHLDREmployment	-15188.145	1389.483	-10.931	0.00000	***
HHLDRWonlyNonHisp	-11922.761	1847.112	-6.455	0.00000	***
HHLDRBonly	-19311.399	1646.219	-11.731	0.00000	***
HHLDROthRace/Origin	-4766.394	1738.842	-2.741	0.00787	**
HHLangEnglish	11365.225	1619.310	7.019	0.00000	***
HHLDRChildren	2603.982	1036.918	2.511	0.01449	*
HHLDRForeignBorn	9503.744	1844.688	5.152	0.00000	***
LogitTractPWonlyNH	-3329.390	272.777	-12.206	0.00000	***

Significance codes: ‘***’ 0.001, ‘**’ 0.01, ‘*’ 0.05, ‘.’ 0.1 ,‘ ’ 1

Note: Income in thousands.

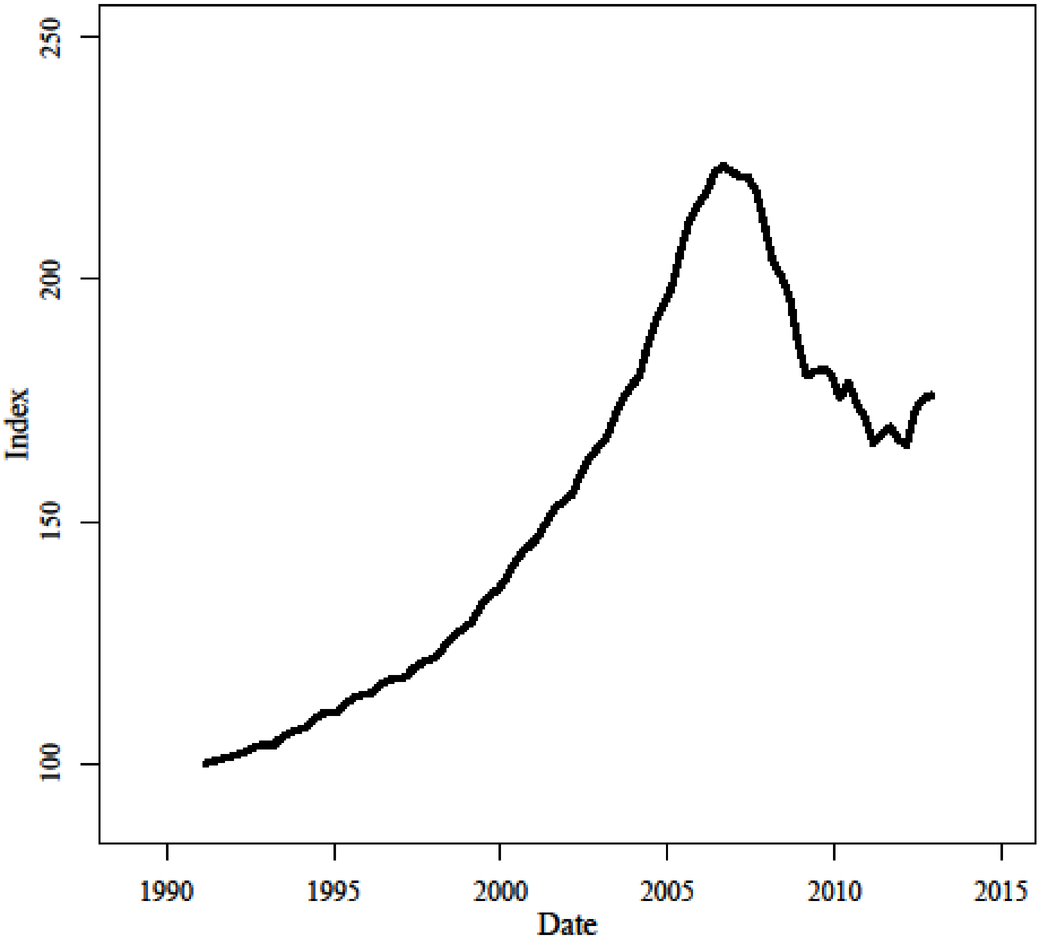
Table 2: Regression of generalized logit relative difference between ACS householder's self-assessed home value and maximum of available CoreLogic county record value items on selected characteristics of the householder, household, and locational context

Variable	Coefficient	Standard Error	t	Pr(>|t|)	significance
(Intercept)	0.52183	0.01798	29.029	0.00000	***
HHIncome	0.00033	0.00003	12.324	0.00000	***
HHLDRAge	-0.00168	0.00022	-7.827	0.00000	***
Residence Duration	0.00590	0.00023	26.036	0.00000	***
Years of Schooling	-0.00305	0.00082	-3.710	0.00043	***
Own Free & Clear	-0.03091	0.00688	-4.491	0.00003	***
HHLDREmployment	-0.01677	0.00643	-2.607	0.01129	*
HHLDRWonlyHisp	-0.03602	0.00933	-3.862	0.00026	***
HHLDRBonly	0.09652	0.00799	12.077	0.00000	***
HHLDROthRace/Origin	-0.02973	0.00883	-3.366	0.00128	**
HHLangEnglish	0.02713	0.00719	3.771	0.00035	***
HHLDRChildren	-0.00871	0.00597	-1.458	0.14952	
HHLDRForeignBorn	-0.01751	0.00798	-2.196	0.03160	*
LogitTractPWonlyNH	0.04713	0.00135	35.014	0.00000	***

Significance codes: ‘***’ 0.001, ‘**’ 0.01, ‘*’ 0.05, ‘.’ 0.1 ,‘ ’ 1

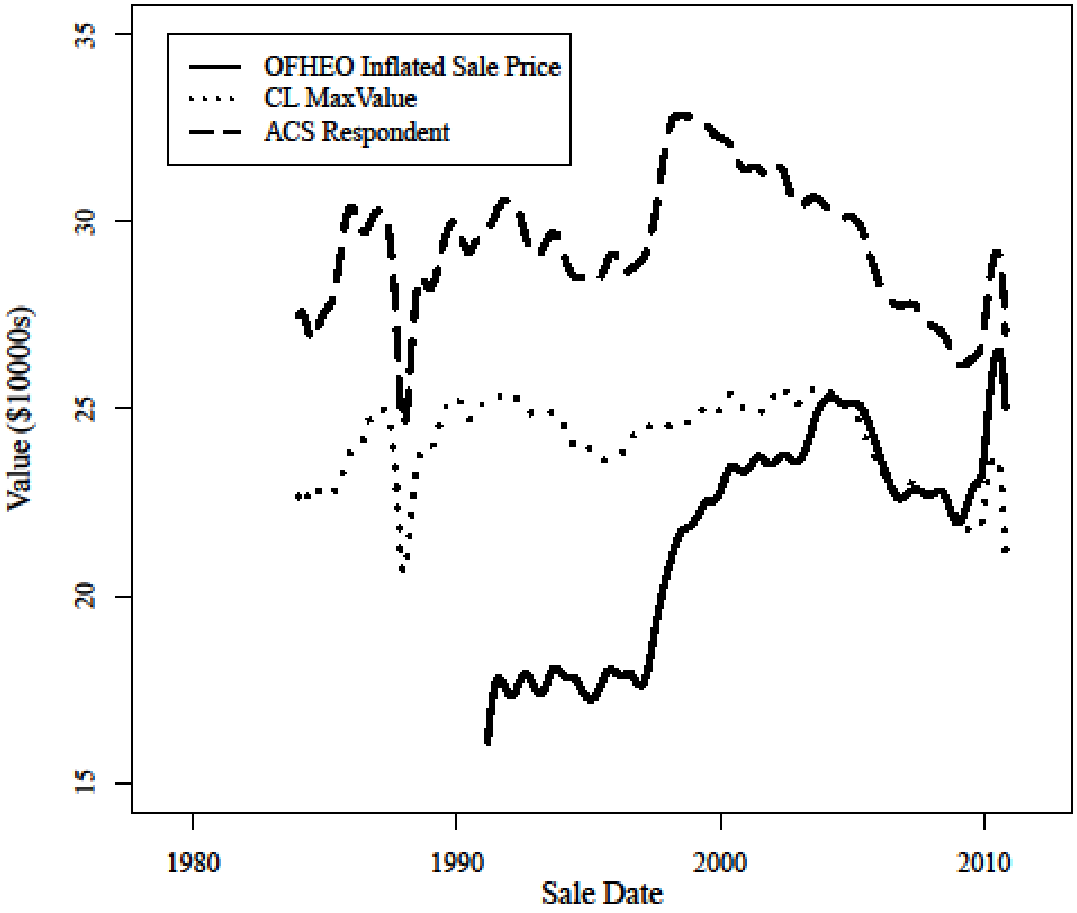
Note: Income in thousands.

Figure 1: US Housing Price Index (OFHEO)



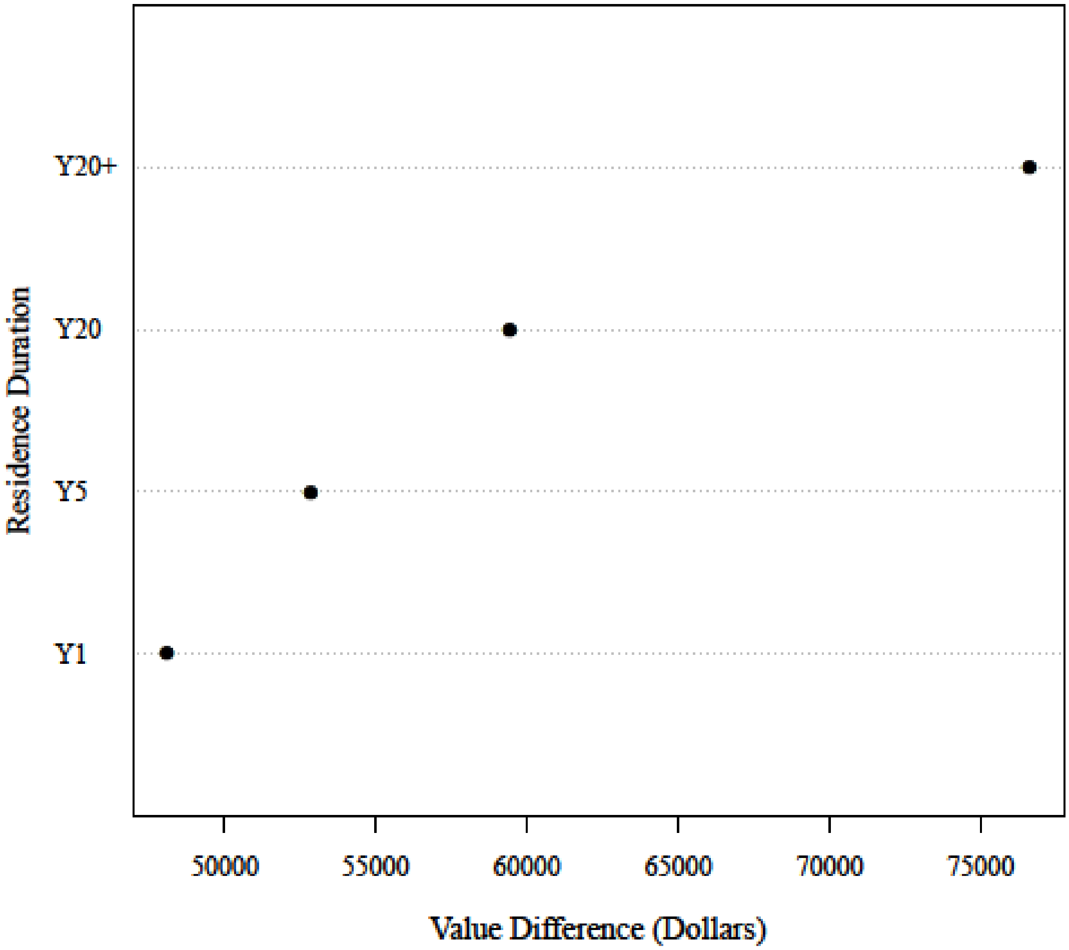
Source: Federal Housing Finance Agency

Figure 2: Smoothed 2009 Housing Value Estimates



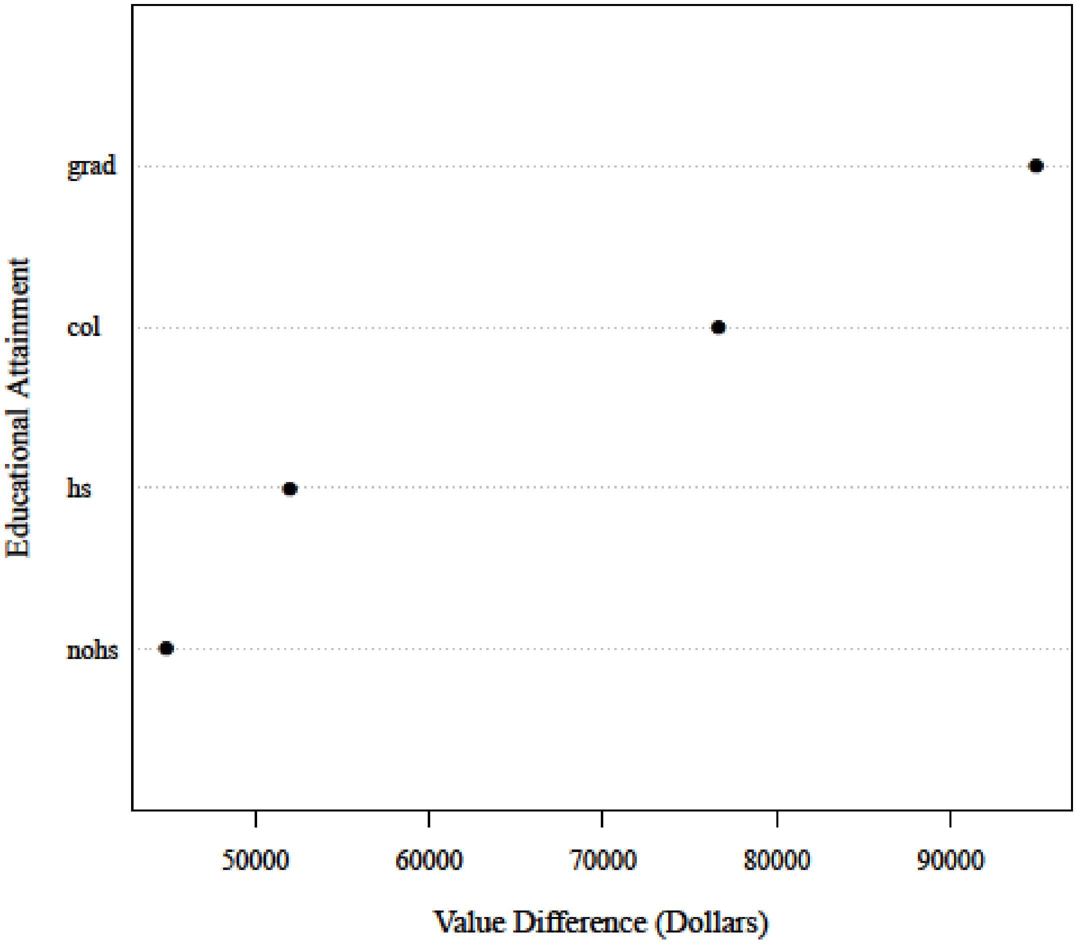
Source: See text.

Figure 3: Mean Difference Housing Value ACS Respondent − CoreLogic



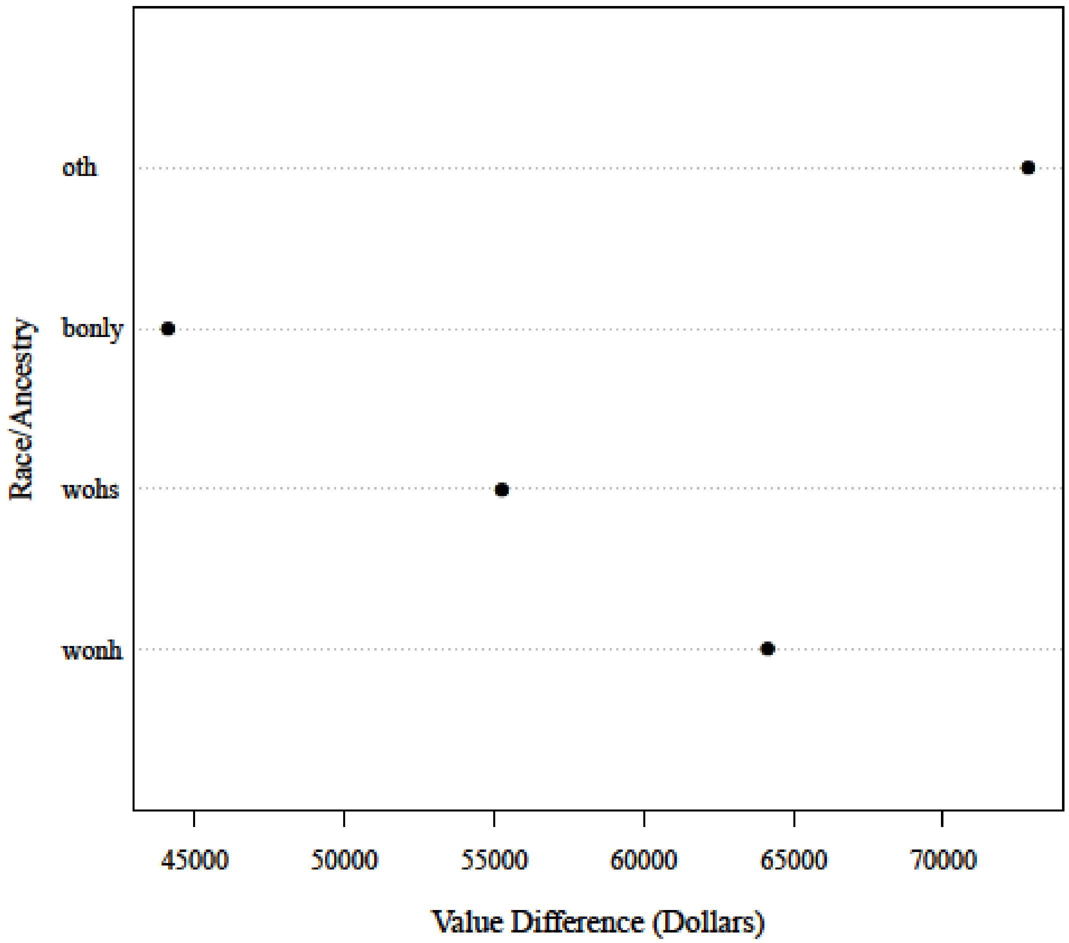
Source: Dataset described in text.

Figure 4: Mean Difference Housing Value ACS Respondent – CoreLogic



Source: Dataset described in text.

Figure 5: Mean Difference Housing Value ACS Respondent – CoreLogic



Source: Dataset described in text.