Data Shop

Data Shop, a department of Cityscape, *presents short papers or notes on the uses of data in housing and urban research. Through this department, PD&R will introduce readers to new and overlooked data sources and to improved techniques in using well-known data. The emphasis will be 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.*

Weighting and the American Housing Survey

Gregory J. Watson
The Moran Company

All the views expressed in this article are the author's own and do not necessarily reflect those of the U.S. Department of Housing or the author's present and past employers.

Abstract

The American Housing Survey (AHS) is the preeminent source of housing characteristic information for the U.S. housing stock. To produce accurate universe-level estimates or other statistics, however, researchers must properly weight the sample observations. This article describes the general strategy used for weighting and then adds notes for researchers who use the AHS.

Introduction

How many housing units are in the Northeast? What percentage of housing units are owner occupied? How many units lack plumbing facilities?

Researchers can answer these questions and many more by using the American Housing Survey (AHS). Answering the questions appropriately requires researchers to understand more than just the housing data collected in the AHS; they must also understand some of the subtleties of how sampling works in the AHS. An understanding of sampling is necessary to understand how the weights are created. These weights enable researchers to produce research with such information as national-level estimates of the number of housing units or the percentage of housing units with particular characteristics.

The goal of this article is to inform researchers about weighting topics in the AHS, including the following:

- The background of the AHS.
- How weights are created.
- Different weight variables.
- Cross-sectional analysis (practical discussion).
- Time-series analysis (practical discussion).
- Special circumstances.

Although the topic of weights is not of the highest importance to housing researchers, it is an essential topic.

Because this article provides an overview about weighting and the AHS, researchers are strongly encouraged to closely read the technical appendixes in the resources listed in the Additional Reading section.

Background on the AHS

The AHS is funded by the U.S. Department of Housing and Urban Development (HUD) and conducted by the Census Bureau. The AHS collects very detailed information on housing units and their occupants; the data originally was collected via paper surveys and now is collected via telephone interviews. The national sample, consisting of about 50,000 housing units, is conducted every 2 years, and the metropolitan sample is conducted on a rotating basis across different metropolitan areas. Although this article focuses on the national sample, the weighting issues addressed herein are mostly the same as those that pertain to the metropolitan samples.

The current national sample was drawn in 1984 and first implemented in 1985. As a matter of design, the original sampling strategy was one of stratified random sampling, with oversampling (sampling a greater proportion) conducted inside certain strata to get better representation. The same housing units (structures) are surveyed in each survey, which enables the tracking of units over time as a large panel data set. The sample is adjusted over time to account for units being removed from and added to the housing stock.

The original sampling strategy, combined with adjustments over time, led to a requirement to use sample weights to produce accurate universe-level estimates or sample proportions. Failure to use the weights will lead to erroneous estimates of counts or proportions, which can lead researchers to erroneous conclusions. The weight assigned to each sample case in the AHS is the number of housing units represented by that particular sample case. Depending on the year, the average weight assigned to each case is in the range of 2,000 to 2,200, representing that number of housing units.

Researchers must also remember that the sampling strategy is set up to produce estimates of the number of housing units, not estimates of population or the number of households. Estimates of population created using the AHS weight should be treated with extreme caution. They may be reasonably accurate, but, because it is hard to know for sure, benchmarking with other data sets is appropriate and highly recommended.

Other data sets to compare against for estimates of both population and housing units include the Current Population Survey (CPS), American Community Survey (ACS), and decennial census. When doing very quick benchmarking, researchers should remember that it is not always necessary to download the microdata and create summary tables. The Census Bureau provides a very useful tool in American FactFinder (http://factfinder.census.gov) that allows for quick access to published statistics from its reports. Fannie Mae Foundation's DataPlace (http://www.dataplace.org) similarly allows for very easy access to summary statistics that are useful for benchmarking.

Weights in the AHS

Depending on the year of the AHS, two or three weights are present in the data files, but researchers commonly use only one.

The first weight is the "pure weight"—the "PWT" variable in the AHS data—which is used as the initial basis for the adjusted household weight. The pure weight is the inverse of the probability of selection based on the original sampling. If the AHS had been implemented with a pure random sample, this value would be the same for every case; however, because stratified random sampling occurs, the values are different depending on the strata. A higher weight means the sample observation represents more units, which means the housing unit had a lower chance of being selected in the first place. This stratified random sampling is the reason why sample proportions should be computed on weighted data instead of on unweighted data. Merely calculating proportions without weighting data may result in computing incorrect proportions.

Generally, the pure weight should not vary over time; however, by design, certain circumstances will occur in which the pure weight does vary. The pure weight will vary when the sample changes. In certain years, the AHS was designed to oversample inside certain groupings. This oversampling provides more sample cases in certain strata that allow for a higher level of statistical confidence in the estimates produced. This oversampling, however, leads to changes in the pure weights for those strata.

In select national surveys conducted from 1985 to 1993, the AHS oversampled in rural areas. In 1995, the AHS started oversampling in the six largest metropolitan areas: Chicago, Detroit, Los Angeles, New York, Northern New Jersey, and Philadelphia. The oversampling has occurred in every other national survey of the AHS since 1995 but is scheduled to be discontinued starting in 2007.

The pure weight is generally not interesting in and of itself but is useful in certain circumstances. The pure weight is used as the basis for the adjusted weight, which is why it is of interest now.

In certain cases, the pure weight inexplicably changes due to historical errors in the data processing. These cases are few and not material.

The household adjusted weight—the "WEIGHT" variable in the AHS data—is where nearly all the researcher's interest should be. This variable originally was based on the pure weight variable discussed previously but then is adjusted by the Census Bureau. This weight is adjusted to control for changes in the sample, such as losses in the housing stock or other adjustments. These adjustments are based on benchmarking with other data sets, such as the CPS and the decennial census.

The adjusted weights should nearly always be used in analysis because they provide the most accurate estimates of the housing stock. These weights do change from year to year, however, so, although the weights are extremely useful for cross-sectional analysis, adjustments need to be made if researchers try to link multiple years of data. These weights change from year to year both because of changes in the sample (just as the pure weight changes because of oversampling) and because of smaller adjustments due to changes in the sample, such as the addition of new construction. These changes are necessary because the sample originally was drawn in 1984 and many changes have occurred since then. When it sets the adjusted weights, the Census Bureau also corrects for the problem of nonresponse from the housing unit occupants.

With the 2001 data, a second adjusted weight variable started being released with the AHS. This variable—WGT90GEO—is based on the 1990 geography definitions, not the 1980 geography definitions that were used when the samples were created. As reported in the AHS codebook for 1997 and successive survey years, "HUD and Census recommend that WGT90GEO, the 1990 geography-based weight, be used only to match numbers from the public use file (PUF) with numbers in the publication at the U.S. and Census region level. For historical comparisons and other analyses, use the 1980 geography-based weights (WEIGHT), as these are comparable to previous publications."

Zero Weights

In a few circumstances, a sample observation will have a zero weight assigned to it. This assignment occurs when the unit is *permanently* removed from the housing stock, which is known as a Type C removal from the housing stock. These units will have a weight of 0 for both the adjusted weight and the pure weight for the last year they are present; other possible reasons for a zero weight include an interview conducted in error or certain other interviews not conducted in housing units. After the last record of the unit's change in status, the case is removed from the sample. In contrast, Type A and B noninterviews will still have a nonzero pure weight. These Type A and B observations remain in the data so researchers can examine the characteristics of units being permanently removed from the housing stock.

Changes in Weights

As discussed previously, both the pure weight and adjusted household weight will change due to changes in the sample, but only the household adjusted weight will be modified from year to year to take into account issues other than oversampling.

Practical Discussion—Cross-Sectional Analysis

From a practical perspective, researchers do not need to remember much to be able to properly apply weights to the AHS; however, they do need to keep the following important items in mind:

1. Use the household adjusted weight variable (WEIGHT).

2. Use the correct values when reading the weight in and then using it. The weights should have two decimal places and have an average of a little more than 2,000, depending on the year of the survey. When reading the data in, be certain about whether the raw data has the weight variable with the decimal places explicitly or implicitly defined. The data in the ASCII versions of the files generally have the implied decimal places, so researchers must divide the weight by 100 to put in the decimal places.

   If analyzing data using SAS, use the WEIGHT option, not the FREQ option, to weight the sample cases. The FREQ option truncates the integer value and removes the decimal places.

3. If you get a warning about a zero weight, check the data but do not be overly concerned about the zero weight.

4. It is possible that a valid housing unit is vacant, which occurs when a unit has been sold or new construction has been completed but the unit has not been occupied yet or when a unit simply was not occupied when the interview occurred. Vacant units still have a valid weight assigned to them. As a result, any analysis of occupied units must be run on a restricted sample. Restrict the analysis to occupied housing units by restricting the analysis to observations in which the status indicator equals 1.

Practical Discussion—Time-Series Analysis

One of the elegant design elements of the AHS is that different years of data can be linked together to perform time-series analysis of the housing units. Comparing the characteristics of a particular unit from one year to the next is a relatively easy analysis, but estimating the number of units this case represents is difficult.

Housing units in different years of the AHS can be linked together using the CONTROL variable. The CONTROL variable, the unique identifier for the housing unit, stays constant from year to year. By comparing a unit's characteristics from year to year, researchers can identify changes in characteristics of the housing unit or its occupants. Given the changes in weight from year to year, researchers face the question of what weight to use.

This section provides an extremely brief and general discussion about how to create a weight for time-series analysis. Researchers are strongly encouraged to refer to the documentation for the Components of Inventory Change (CINCH) reports at http://www.huduser.org/datasets/cinch.html for more detail. A warning to researchers: computing these new weights for your purposes is not a simple or easy task. In addition, you should use caution when viewing these weights.

A problem with linking data across years is that the estimated number of units represented by the observation will vary from year to year. That estimate does also not change consistently from one unit to another, so it is not simply a case of "inflation." The weight will vary simply because of the adjustments made by the Census Bureau; the Census Bureau is controlling for other changes it has measured in the housing stock. In addition, much larger adjustments will occur due to changes in the oversampling.

Researchers need to limit the sample from both years to just those sample cases that appear in both years. In other words, researchers need to exclude the cases that are in the AHS due to oversampling, precisely because they do not appear in both years. Then, after the sample is limited to those cases that appear in both years, the weights need to be adjusted to take those excluded cases into account. As part of this step, however, researchers must be careful not to improperly exclude known added units (for example, new construction) or known removed units (for example, those lost to a disaster) because these units appropriately should be in only 1 year of the data.

As a general rule, the weight then can be set to the maximum of the individual weights. Using the maximum gets very close to the new weight that needs to be used, but some refinements still need to be done. The next refinement is to account for cases that legitimately appear in only one year, such as units that were added or removed from the housing stock. For the weights for these special cases, use the weight that is present in the single year that the unit exists as an occupied unit.

After the weights are approximately set, then "ratio adjustment" should be performed to more precisely set the weights to match to published control totals. The ratio adjustment process is only performed on those cases present in both years. Excluded are the weights of cases that are present in only 1 year, namely the known additions and known removals.

The process of ratio adjustment roughly consists of summing up all the current weights, computing the ratio of that total to the published control total, and then applying that ratio to all the individual weights to create a new, "ratio-adjusted" weight for each case. Other analyses can then be computed using these new computed weights.

Special Cases of Weights Changing Dramatically

In certain situations, the sample weight assigned to a housing unit can change dramatically. Some possibilities include Type C interview losses, conversions/mergers, and other corrections made by the Census Bureau. These situations generally represent a very small proportion of the cases and are generally not issues when working with a sufficiently large sample size. A Type C loss has a weight of 0, and the case generally is removed from the sample in the following year because no possibility exists of the unit returning to the housing stock. In comparison, Type A and B noninterviews maintain their weights.

Small Sample Size Caveats

This article has thus far presented reasons for weighting the data and notes about weighting the data. Even with weights, however, estimates derived from the data may not be perfect because the

AHS is a survey and although the weights are intended to allow estimates, there is still the potential for measurement error. Therefore, researchers must be careful, especially when using small sample sizes, to recognize that potentially erroneous estimates exist due to sampling error.

Researchers are strongly encouraged to read "Sampling Errors for Small Groups," available at http://www.huduser.org/datasets/ahs/ahsprev.html, for a more detailed discussion about issues related to small sample sizes and the AHS. Researchers are also encouraged to read appendix D in any AHS report; this appendix is also available on line at http://www.census.gov/hhes/www/housing/ahs/errors.pdf.

Generally, the smaller the number of sample cases used to create the total estimates, the wider the confidence intervals surrounding the measurement. Researchers should pay particular attention to this issue when comparing different estimates. Because each case has a sample weight of approximately 2,000 housing units, it is highly unlikely that researchers would be able to obtain reliable estimates of anything less than several thousand units.

Conclusion

It is hoped that this brief discussion of AHS weights will encourage the appropriate use of weights. Using appropriate weights will lead to more accurate estimates of the characteristics of the U.S. housing stock.

Acknowledgments

The author thanks all the staff at the U.S. Department of Housing and Urban Development and the Census Bureau that he and his colleagues at ICF Consulting (now ICF International) had the pleasure to work with on American Housing Survey issues in the past. These individuals include, but are not limited to, Tom Blatt, Fred Eggers, Paul Harple, Jane Kneessi, Dennis Schwanz, Ron Sepanik, David Vandenbroucke, and Barbara Williams.

Author

Gregory J. Watson is a partner at The Moran Company, a healthcare consulting firm.

Additional Readings

For additional details, researchers are strongly encouraged to refer to the following documents:

Hadden, Louise, and Mireille Leger. 1995. *Codebook for the American Housing Survey, Volume 1*. Report prepared by Abt Associates, Inc. Washington, DC: U.S. Department of Housing and Urban Development. http://www.huduser.org/datasets/ahs/ahs_codebook.html.

ICF Consulting. 2001. *Documentation of Changes in the 1997 American Housing Survey*. Report prepared for U.S. Department of Housing and Urban Development, Office of Policy Development and Research. http://www.huduser.org/intercept.asp?loc=/Datasets/ahs/docchg1997.pdf.

ICF International. 2006. *Codebook for the American Housing Survey, Public Use File: 1997 and Later.* Report prepared for U.S. Department of Housing and Urban Development, Office of Policy Development and Research. http://www.huduser.org/intercept.asp?loc=/Datasets/ahs/AHS_Codebook.pdf

U.S. Department of Housing and Urban Development. 1998. *Sampling Errors for Small Groups.* http://www.huduser.org/datasets/ahs/binom.exe.

U.S. Department of Housing and Urban Development, contracting with various firms. Multiple years. *Components of Inventory Change (CINCH)* reports. http://www.huduser.org/datasets/cinch.html.

U.S. Department of Housing and Urban
Development
Office of Policy Development and Research
Washington, DC 20410-6000

Contents

U.S. DEPARTMENT OF HOUSING AND URBAN DEVELOPMENT

Cityscape

VOLUME 9, NUMBER 2 • 2007