Independent Statistics & Analysis
U.S. Energy Information Administration

Residential Energy Consumption Survey (RECS):

Using the 2015 microdata file to compute estimates and standard errors (RSEs)

May 2017

Independent Statistics & Analysis
www.eia.gov

U.S. Department of Energy
Washington, DC 20585

This report was prepared by the U.S. Energy Information Administration (EIA), the statistical and analytical agency within the U.S. Department of Energy. By law, EIA's data, analyses, and forecasts are independent of approval by any other officer or employee of the United States Government. The views in this report therefore should not be construed as representing those of the U.S. Department of Energy or other federal agencies.

Table of Contents

Overview

EIA makes available a public-use microdata file for each RECS survey cycle. The 2015 file is a valuable tool for users conducting detailed analysis of home energy use. This document provides some background on the RECS design, as well as useful tips and examples that will guide users through the use of the RECS microdata.

RECS sample design

The RECS sample was designed to estimate energy characteristics, consumption, and expenditures for the national stock of occupied housing units and the households that live in them. The 2015 RECS allows for separate estimation for Census regions and divisions. (The return to the traditional sample size for the 2015 RECS does not allow for state-level estimation, as was available for the expanded 2009 RECS.) To produce estimates for these geographies and the total U.S., the sample cases were properly weighted to represent the population, including the residences not in the sample. In a sense, a case's weight indicates the number of households that the particular case represents.

Base sampling weights, which are the reciprocal of the probability of being selected for the RECS sample, were first calculated for each sampled housing unit. The base weights were adjusted to account for survey nonresponse and ratio adjustments were used to ensure that the RECS weights add up to Census Bureau estimates of the number of occupied housing units for 2015. The variable **NWEIGHT** in the data file represents the *final sampling weight*, accounting for different probabilities of selection and rates of response, and being adjusted for the Census Bureau housing unit estimates. NWEIGHT is the number of households in the population that the observation represents. For example, if NWEIGHT for a household is 10,000, that household represents itself and 9,999 other non-sampled households. More details about the sample design can be found in the *RECS 2015 Technical Documentation – Summary*.

Sampling error

Estimates from a sample survey like RECS are not exact but are statistical estimates with some associated sampling error in each direction—the result of generating estimates based on a sample rather than a census of the entire population. Sampling error provides a measure of the accuracy of a particular estimate for a characteristic based on how common and variable it is in the population, given a particular sample size.

Standard errors are used in conjunction with survey estimates to measure sampling error, construct confidence intervals, or perform hypothesis tests. A relative standard error (RSE) is defined as the standard error (square root of the variance) of a survey estimate, divided by the survey estimate, and multiplied by 100. In other words, the RSE is the standard error relative to the survey estimate on a scale from zero to 100. The larger the RSE, the less precise the survey estimate is of the true value in the population. An RSE is shown for each estimate in the RECS tables.

Fay's balanced repeated replication (BRR) method of estimating standard error

RECS uses Fay's method of the balanced repeated replication (BRR) technique for estimating standard errors. This method uses replicate weights to repeatedly estimate the statistic of interest and calculate the differences between these estimates and the full-sample estimate.

See Fay (1989), Heeringa, West, and Berglund (2010), Judkins (1990), Lee and Forthofer (2006), Roa and Shao (1999), Rust (1985), and Wolter (2007) for technical details.

If θ is a population parameter of interest, let $\hat{\theta}$ be the estimate from the full sample for θ. Let $\widehat{\theta_r}$ be the estimate from the r-th replicate subsample by using replicate weights and let ε be the Fay coefficient, 0 ≤ ε < 1. The variance of $\hat{\theta}$ is estimated by:

$$\hat{V}(\tilde{\theta}) = \frac{1}{R(1-\ \varepsilon)^2}\sum_{r=1}^{R}(\widehat{\theta_r} - \hat{\theta})^2$$

For the 2015 RECS, R=96 (the number of replicate subsamples) and ε = 0.5. The formula for calculating the RSE is:

$$\left(\frac{\sqrt{\hat{V}(\hat{\theta})}}{\hat{\theta}}\right) X\ 100$$

Examples: Using final weights (NWEIGHT) and replicate weights to calculate estimates and RSEs

The following instructions are examples for calculating any RECS estimate using the final weights (NWEIGHT) and the associated RSE using the replicate weights. These examples calculate estimates and standard errors about **households using natural gas as their main heating source**, shown in Table HC6.1.

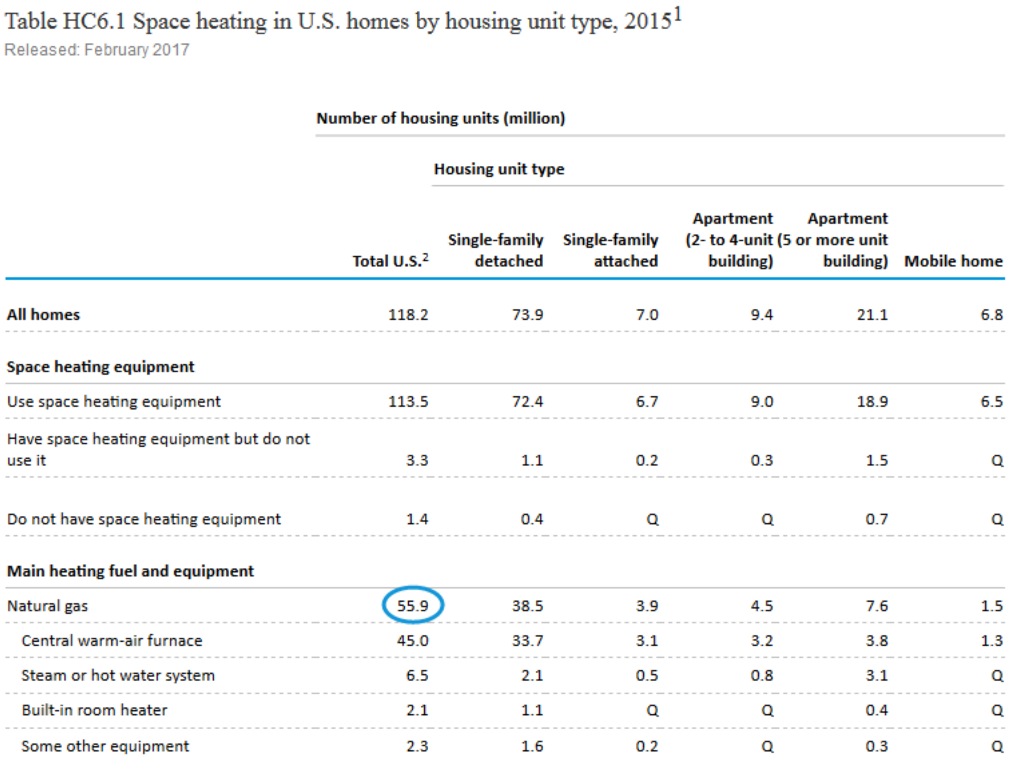

Table HC6.1 Space heating in U.S. homes by housing unit type, 2015[1]

Released: February 2017

	Number of housing units (million)					
		Housing unit type				
	Total U.S.[2]	Single-family detached	Single-family attached	Apartment (2- to 4-unit building)	Apartment (5 or more unit building)	Mobile home
All homes	118.2	73.9	7.0	9.4	21.1	6.8
Space heating equipment						
Use space heating equipment	113.5	72.4	6.7	9.0	18.9	6.5
Have space heating equipment but do not use it	3.3	1.1	0.2	0.3	1.5	Q
Do not have space heating equipment	1.4	0.4	Q	Q	0.7	Q
Main heating fuel and equipment						
Natural gas	55.9	38.5	3.9	4.5	7.6	1.5
Central warm-air furnace	45.0	33.7	3.1	3.2	3.8	1.3
Steam or hot water system	6.5	2.1	0.5	0.8	3.1	Q
Built-in room heater	2.1	1.1	Q	Q	0.4	Q
Some other equipment	2.3	1.6	0.2	Q	0.3	Q

We have provided instructions for Excel users and users with access to statistical software. Software packages like SAS/STAT, R, Stata, SUDAAN, and WesVar can process replicate weights to calculate RSEs. Note that while EXCEL can be used to calculate point estimates, it cannot process replicate weights to calculate RSEs for RECS or other complex sample designs with varying probabilities of selection. EIA recommends calculating standard errors or RSEs in conjunction with estimates to account for sampling error.

For Excel Users

Excel Example 1: Calculate the frequency of households that used natural gas as their main space heating fuel

A simple count of households can be estimated using the sum of NWEIGHTS for a specified subset of cases within the RECS data file. For this example, filter the file for all cases where natural gas space heating was used as the main heating fuel (FUELHEAT= 1). There are 2,560 cases with FUELHEAT = 1. By adding the NWEIGHT column for these 2,560 cases, the estimated number of households that used natural gas as main heating fuel was approximately 55,930,144. This is equal to 47% of all homes, or 55.9 million/118.2million (the sum of NWEIGHT for all cases in RECS.)

For SAS Users

SAS Example 1: Calculate the frequency and RSE of households that used natural gas as their main space heating fuel

Create a new variable to flag the records we are interested in - households that used natural gas as their main space heating fuel. This new variable NG_MAINSPACEHEAT is equal to 1 if the household used natural gas as their main space heating fuel, and 0 otherwise.

```
DATA RECS15;
      SET RECS2015_PUBLIC_V1;
      IF FUELHEAT=1 THEN NG_MAINSPACEHEAT =1;
      ELSE NG_MAINSPACEHEAT =0;
RUN;
```

Use the variable NWEIGHT in the WEIGHT statement and the variable NG_MAINSPACEHEAT in the TABLES statement in PROC SURVEYFREQ. To get the sampling error (RSE) associated with the estimate, we can use PROC SURVEYFREQ to process the replicate weights.

```
PROC SURVEYFREQ DATA=RECS15 VARMETHOD=BRR(FAY);
      REPWEIGHTS BRRWT1-BRRWT96;
      WEIGHT NWEIGHT;
      TABLES NG_MAINSPACEHEAT;
RUN;
```

The estimated number of households that used natural gas as their main space heating fuel is 55,930,144. The standard deviation of the frequency is 155,229 and the calculation for the RSE is: (155,229 / 55,930,144)*100 = 0.3. This means that the sampling error is about 0.3% of the estimate, relatively small.

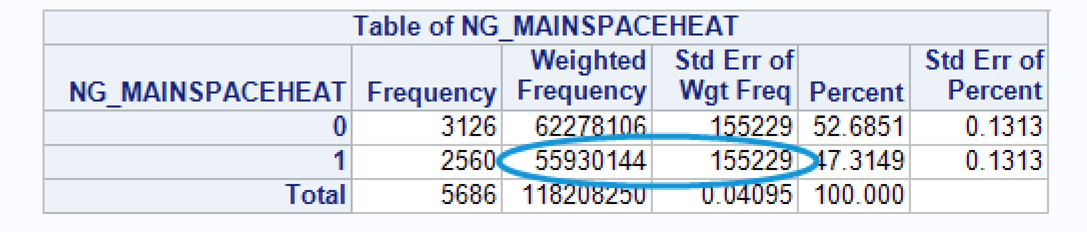

Table of NG_MAINSPACEHEAT					
NG_MAINSPACEHEAT	Frequency	Weighted Frequency	Std Err of Wgt Freq	Percent	Std Err of Percent
0	3126	62278106	155229	52.6851	0.1313
1	2560	55930144	155229	47.3149	0.1313
Total	5686	118208250	0.04095	100.000	

Notes to consider when using the microdata file and replicate weights

1. *Publication standards:* EIA does not publish RECS estimates where the RSE is higher than 50 or the count of households used for the calculation is less than 10 (indicated by a “Q” in the data tables). These are EIA’s recommended guidelines for custom analysis using the public use microdata file.

2. *Imputation variables:* Most variables were imputed for “Don’t Know” and “Refuse” responses. The "Z variables", also referred to as "imputation flags", are included in the public use microdata file. The imputation flag indicates whether the corresponding non-Z variable was based upon reported data (Z variable = 0) or was imputed (Z variable = 1). There are no corresponding "Z variables" for variables from the RECS questionnaire that were not imputed, variables where there was no missing data, and variables that are not from the questionnaire. EIA recommends using the imputed data, where available, to avoid biased estimation.

3. *Standardized coding:* Variables that were not imputed use the response codes -9 for “Don’t Know” and -8 for “Refuse”. Variables that are not asked of all respondents use the response code -2 for “Not Applicable”. For example, if a respondent said they did not use any televisions at home (TVCOLOR = 0) then they were not asked what size of television is most used at home, thus TVSIZE1 = -2. Use caution when performing calculations on variables that may have -2, -8, or -9 responses.

4. *Indicator variables:* The microdata file contains variables to indicate the use of major fuels and specific end uses within each housing unit for 2015. These variables are derived from answers given by each respondent and indicate whether the respondent had access to and actually used the fuel and engaged in the end-use. All indicators are either a 0 or 1 for each combination of major fuel and end-use. For example, a respondent who says they heated their home with electricity in 2015 will have the derived variable ELWARM = 1. If a respondent says they have equipment but did not use it the corresponding indicator will be 0. As an example, a respondent in a cool climate might have air-conditioning equipment but did not use it in 2015. For this case, ELCOOL would be 0.

5. *Confidentiality:* The 2015 RECS was collected under the authority of the Confidential Information Protection and Statistical Efficiency Act (CIPSEA). EIA, project staff and its contractors and agents are personally accountable for protecting the identity of individual respondents. The following steps were taken to avoid disclosure of personally identifiable information on the public use microdata file.

   - Local geographic identifiers of sampled housing units, such as zip codes, were removed.

   - Building America Climate Regions with few sample cases (“Very Cold” and “Mixed- Dry”) were combined with the most similar region.

   - The variable indicating on-site wind generation (WIND) was removed due to too few responses.

- The variable HHAGE (age of the householder) was top-coded at 85.

- Weather and climate (HDD and CDD) values were inoculated with random errors. Adjustments were minor and will not result in significant differences than those estimates displayed in data tables.

References

Fay, R. E. (1989), "Theory and Application of Replicate Weighting for Variance Calculations," in Proceedings of the Survey Research Methods Section, 212–217, American Statistical Association.

Heeringa, S., West, B. T, & Berglund, P. A. (2010). *Applied survey data analysis*. Boca Raton, Fla.: CRC Press.

Judkins, D. R. (1990), "Fay's Method for Variance Estimation," Journal of Official Statistics, 6(3), 223–239.Lee, E. Sul, & Forthofer, R. N. (2006). *Analyzing complex survey data.* 2nd ed.

Thousand Oaks, Calif.: Sage Publications.

Rao, J. N. K. and Shao, J. (1999), "Modified Balanced Repeated Replication for Complex Survey Data," Biometrika, 86(2), 403–415.

Rust, K. (1985), "Variance Estimation for Complex Estimators in Sample Surveys," Journal of Official Statistics, 1(4), 381–397.

Wolter, K. M. (2007). *Introduction to Variance Estimation*, 2nd ed. Springer, New York.

The code and output for this paper was generated using SAS/STAT software, Version 7.11 of the SAS Enterprise Guide for UNIX. Copyright © 2015 SAS Institute Inc. SAS and all other SAS Institute Inc. product or service names are registered trademarks or trademarks of SAS Institute Inc., Cary, NC, USA.