Independent Statistics & Analysis
U.S. Energy Information Administration

Residential Energy Consumption Survey (RECS)

Using the 2009 microdata file to compute estimates and standard errors (RSEs)

February 2013

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 Department of Energy or other Federal agencies.

Table of Contents

Overview *1*

Examples: Using final weights (NWEIGHT) and replicate weights to calculate estimates and RSEs *2*

For Excel users *2*

For SAS users *3*

Notes to consider when using the microdata file and replicate weights *5*

References *7*

Overview

EIA makes available a public-use microdata file for each RECS survey cycle. The 2009 file, which contains over 900 variables, 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 2009 RECS has an increased sample size that allows for separate estimation for Census Regions, Census Divisions, 16 individual states, and remaining groups of states. To produce estimates for these geographies, 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 size of the population that the particular case represents. **Base sampling weights**, which are the reciprocal of the probability of being selected for the RECS sample, were 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 housing units for 2009. 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.

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 $\hat{\theta}_r$ be the estimate from the r-th replicate subsample by using replicate weights and let ε be the Fay coefficient, $0 \leq \varepsilon < 1$. The variance of $\hat{\theta}$ is estimated by:

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

For the 2009 RECS, R=244 (the number of replicate subsamples) and ε = .5. The formula for calculating the RSE is:

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

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**.

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 5,903 cases with FUELHEAT = 1. By adding the NWEIGHT data for these 5,903 cases, the estimated number of households that used natural gas as main heating fuel was approximately 55,622,460. This is equal to 49% of all homes, or 55.6 million/113.6 million (the sum of NWEIGHT for all cases in RECS.)

Excel Example 2: Calculate total and average natural gas space heating consumption for households that used natural gas as their main space heating fuel

To calculate total natural gas space heating consumption, in Btu, for the 55.6 million households who used natural gas as their main heating fuel, multiply NWEIGHT * BTUNGSPH * 1000 (Btu estimates are in 1000s) for each case. Households that did not use natural gas for space heating will have BTUNGSPH = 0, so no recoding or filtering of data is needed. The sum of each weighted total is 2,901,633,736,365,000 (2.9 quadrillion Btu). To calculate the average consumption for these household, divide 2.9 quadrillion Btu by 55.6 million households to equal 52.2 million Btu per household.

For SAS Users

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

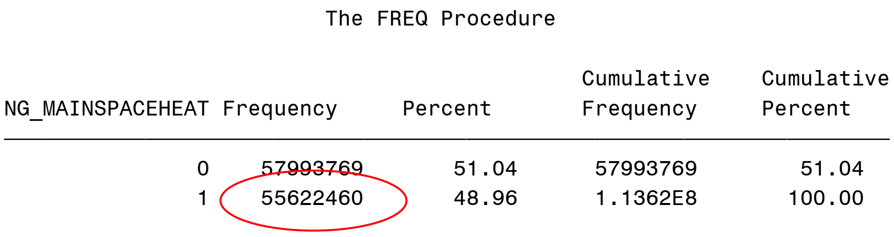
To allow for later RSE calculations, first, merge the replicate weight file with the public-use microdata set by the variable DOEID. Create a new variable to flag the records we are interested in - households that used natural gas as their main space heating fuel. This new variable NG_MAINSPACEHEAT is equal to 1 if the household used natural gas as their main space heating fuel, and 0 otherwise.

```
Data recs09;
      merge recs2009_public_repweights recs2009_public_v3;
      by DOEID;
      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 TABLE statement in PROC FREQ:

```
Proc freq data=recs09;
      weight NWEIGHT;
      table NG_MAINSPACEHEAT;
run;
```

The estimated number of households that used natural gas as their main space heating fuel is 55,622,460.



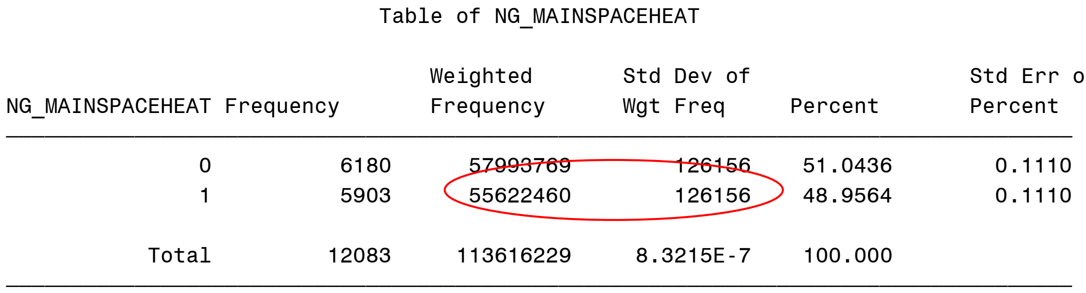
```
                         The FREQ Procedure

                                                   Cumulative    Cumulative
NG_MAINSPACEHEAT  Frequency     Percent            Frequency     Percent
_____________________________________________________________________________

               0     57993769       51.04           57993769         51.04
               1     55622460       48.96           1.1362E8        100.00
```

SAS Example 2: Calculate the RSE for the number of households that used natural gas as their main space heating fuel

To get the sampling error (RSE) associated with the above estimate (55.6 million households), we can use PROC SURVEYFREQ to process the replicate weights.

```
Proc surveyfreq data=recs09 VARMETHOD=BRR(fay);
      repweights brr_weight_1 - brr_weight_244;
      weight NWEIGHT;
      tables  NG_MAINSPACEHEAT;
run;
```

The standard deviation of the frequency is 126,156 and the calculation for the RSE is: (126,156 / 55,622,460)*100 = .2. This means that the sampling error is about .2% of the estimate, relatively small.



```
                              Table of NG_MAINSPACEHEAT

                                  Weighted        Std Dev of                   Std Err of
NG_MAINSPACEHEAT Frequency        Frequency       Wgt Freq      Percent        Percent
____________________________________________________________________________________
               0         6180       57993769        126156      51.0436        0.1110
               1         5903       55622460        126156      48.9564        0.1110

           Total        12083      113616229     8.3215E-7      100.000
____________________________________________________________________________________
```

SAS Example 3: Calculate total and average natural gas space heating consumption by region, and associated RSEs, for households that used natural gas as their main space heating fuel

To calculate total and average consumption for a specified subset of households in SAS use the SURVEYMEANS procedure. For this example, use BTUNGSPH in the VAR statement, and the newly created variable NG_MAINSPACEHEAT in the DOMAIN statement. For a further breakout of consumption, add a second dimension to the DOMAIN statement. For this example, Census region (REGIONC) is added. The WEIGHT and REPWEIGHT variables are the same as the PROC SURVEYFREQ example above. Use the options *sum, clsum, mean*, and *clm* to request the sum, confidence interval for the sum, mean, and confidence limit of the mean, respectively, of the variable BTUNGSPH. Note that this code calculates an average natural gas space heating consumption, by region, only for those households who used natural gas as their main space heating fuel.

```
Proc surveymeans data=recs09 varmethod=BRR(fay) mean clm sum clsum;
      repweights brr_weight_1 - brr_weight_244;
      weight NWEIGHT;
      domain NG_MAINSPACEHEAT * REGIONC;
      var BTUNGSPH;
run;
```

The first table of output shows the total natural gas consumption by region (in thousand Btu) the standard deviation (error), and the 95% confidence interval. (Note that the estimates for NG_MAINSPACEHEAT = 0 reflect consumption for homes that use a non-natural gas fuel as the main space heating fuel. This is secondary heating consumption for these homes.)

The RSE for the total in the Northeast (REGIONC = 1) is .016 quadrillion Btu/.680 quadrillion Btu*100 = 2.4%. The lower 95% confidence limit is .649 quadrillion Btu and the upper 95% confidence limit is .710 quadrillion Btu. This means that if the sample were repeatedly taken and the confidence interval were constructed from each sample, then 95% of the time those confidence intervals would cover the true population mean.

```
                    Domain Analysis: NG_MAINSPACEHEAT*Census Region

                  Census
NG_MAINSPACEHEAT   Region  Variable     Sum              Std Dev          95% CL for Sum
__________________________________________________________________________________________

               0        1  BTUNGSPH    7935786596       2293558324      3418086801 1.24535E10
                        2  BTUNGSPH   10190146390       2764231755      4745345140 1.56349E10
                        3  BTUNGSPH   16840886141       2767647062      1.13894E10 2.22924E10
                        4  BTUNGSPH    7761940075       1293363575      5214357878 1.03095E10
               1        1  BTUNGSPH  680068469654      15666939273      6.49209E11 7.10928E11
                        2  BTUNGSPH  1.2083405E12      18628130528      1.17165E12 1.24503E12
                        3  BTUNGSPH  527387974140      14153446653      4.99509E11 5.55266E11
                        4  BTUNGSPH  485827202894      21046560187      4.44371E11 5.27283E11
```

The second table of output below shows the average consumption, the standard error, and 95% confidence limits. For the Northeast, this results in an average of 63.0 million Btu per household, an RSE of 2.1%, and a 95% confidence interval of 60.4 million Btu to 65.6 million Btu.

```
                                The SURVEYMEANS Procedure

                      Domain Analysis: NG_MAINSPACEHEAT*Census Region

                  Census                                 Std Error
NG_MAINSPACEHEAT  Region  Variable     Mean              of Mean           95% CL for Mean
__________________________________________________________________________________________

               0        1  BTUNGSPH     795.098906       228.949742       344.1288  1246.0690
                        2  BTUNGSPH    1267.786764       344.652301       588.9134  1946.6601
                        3  BTUNGSPH     586.098340        96.197367       396.6151   775.5816
                        4  BTUNGSPH     690.487078       114.378481       465.1919   915.7823
               1        1  BTUNGSPH          63030      1311.878583     60446.1379 65614.2416
                        2  BTUNGSPH          67581      1011.293003     65589.4360 69573.3924
                        3  BTUNGSPH          39516      1052.972299     37441.7019 41589.8526
                        4  BTUNGSPH          35705      1481.069216     32787.2825 38621.9073
__________________________________________________________________________________________
```

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 computers at home (COMPUTER = 0) then they were not asked what type of computer is most used at home, thus PCTYPE1 = -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 2009. 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 2009 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 2009. For this case, ELCOOL would be 0.
5. *Data quality variables:* Data collected from energy suppliers is sometimes incomplete or excludes consumption or cost for particular end uses. Reported data also may need to be adjusted because it includes data for non-household uses or data for other housing units. A series of consumption and expenditures data quality variables are included at the end of the microdata file to inform data users of the origin and quality of the energy supplier data used to calculate fuel totals for each sampled case.
6. *Confidentiality:* The 2009 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 year of construction for sampled housing units (YEARMADE) was bottom coded at 1920.
   - Two variables were masked to prevent identification of large multi-family residential buildings sampled in 2009: NUMFLRS (number of floors in a 5+ unit apartment building) and NUMAPTS (number of apartments in a 5+ unit apartment building). Households with NUMFLRS greater than 15 were replaced with the mean of the values above 15 by Census region. To give a very simple example, if there were only three households in a Census region with NUMFLRS greater than 15 with NUMFLRS values of 20, 25, and 30, then the NUMFLRS values for all three would

be 25. Similarly, households with NUMAPTS greater than 200 were replaced with the mean of the values above 200 by Census region.

- The variable indicating the type of on-site electricity generation (ONSITETYPE) was removed due to too few responses.
- The variable HHAGE (age of the householder) was top-coded at 85.
- Household member ages other than the householder (AGEHHMEM2-14) were categorized.
- 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.
- Consumption and expenditures 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 9.2 of the SAS System for Windows. Copyright © 2002-2008 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.