Independent Statistics & Analysis
U.S. Energy Information Administration

2020 Residential Energy Consumption Survey: Using the microdata file to compute estimates and relative standard errors (RSEs)

June 2022

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 a public-use microdata file available for each *Residential Energy Consumption Survey* (RECS) survey cycle. The 2020 file, available in both SAS and CSV formats, allows users to conduct detailed analysis of home energy characteristics. This document provides some background on the RECS design, as well as useful tips and examples that will help users use the RECS microdata.

Because the sample was not designed to estimate all survey variables at the state level, some estimates may not be reliable due to insufficient sample size. Please use discretion when interpreting results from the microdata.

RECS sample design

EIA designed the RECS sample to estimate energy characteristics, consumption, and expenditures for the national stock of occupied housing units and the people who live in them. For the 2020 RECS, in addition to the ability to estimate household characteristics and energy use for census regions and divisions, EIA added the ability to estimate at the state level for all 50 states and the District of Columbia (DC). This feature was not available in previous survey cycles. In 2015, RECS was not designed to make any state-level estimates, and in 2009 and preceding cycles, estimates were only available at the state level for more populous states. To produce estimates for states, divisions, regions, and the total United States in the 2020 RECS, EIA weighted the sampled housing units to represent the total in-scope population. In a sense, a housing unit's weight indicates the number of housing units that the particular case represents.

As part of the weighting process, EIA first calculated base sampling weights, which are the reciprocal of the probability of being selected for the RECS sample, for each sampled housing unit. EIA then adjusted the base weights to account for survey nonresponse and eligibility. In addition, EIA used poststratification adjustments to ensure that the RECS weights add up to U.S. Census Bureau estimates of the state-level number of occupied housing units for 2020. The variable NWEIGHT in the data file represents the final sampling weight, accounting for different probabilities of selection, rates of response, and adjustment for the U.S. 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 and weighting adjustments are available in the *2020 RECS Household Characteristics Technical Documentation Summary*.

Sampling error and relative standard error (RSE)

Estimates from a sample survey like RECS are subject to sampling error, which occurs because estimates are based on a sample rather than a census of the entire population.

Standard errors are used in conjunction with survey estimates to measure relative amounts of sampling error, construct confidence intervals, or perform hypothesis tests. Similar to previous RECS, the 2020 RECS data tables present RSEs. An RSE is formulated as the standard error (square root of the sampling variance) of a survey estimate, divided by the survey estimate, and multiplied by 100. In other words, the RSE quantifies how much the measured responses vary compared with an average response, expressed as a percentage. The smaller the RSE, the more precise a survey estimate is in terms of its variation around the average value over the replicates. An RSE is shown for each estimate in the RECS tables on a separate tab in the table. Estimates greater than zero

with a corresponding RSE of 0.00 indicate a variable was used as a control total in poststratification. Instructions for calculating RSEs for microdata analysis in SAS and R statistical software are below.

Jackknife method of estimating standard error

The 2020 RECS uses the Jackknife method to produce replicate weights for the responding sample to calculate standard errors of an estimate of interest. This method uses replicate weights to repeatedly estimate the statistic of interest from each of multiple replicate samples generated from the full sample and calculates the differences between these estimates and the full-sample estimate. EIA constructed 60 jackknife replicates to produce variance estimates for univariate statistics with 59 nominal degrees of freedom. The mathematical formula for the variance estimation is expressed below (See Lohr, S.L. (2010) for more technical details).

If θ is a population-weighted estimator, let $\hat{\theta}$ be the estimate from the full sample for θ. Let $\hat{\theta}r$ be the estimate for the r-th replicate, and R is the total number of the replicate weights, the variance of $\hat{\theta}$ is estimated by:

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

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 (NWEIGHT1 – NWEIGHT60). EIA has provided instructions for Excel users and users with access to SAS/STAT and R. Software packages such as SAS/STAT, R, Stata, SUDAAN, and WesVar can process replicate weights to calculate RSEs. Although 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 (estimates only, no RSEs)

Excel Example 1: Calculate the frequency of households that used natural gas as their main space heating fuel (Table HC6.1)

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:

Step 1. Filter the file for all cases where natural gas space heating was used as the main heating fuel (FUELHEAT= 1). There are 8,615 cases with FUELHEAT = 1.

Step 2. Add the NWEIGHT column for these 8,615 cases.

Answer: The estimated number of households that used natural gas as main heating fuel was approximately 56,245,388 households. This amount is equal to 46% of all homes, or 56.25 million / 123.53 million (the sum of NWEIGHT for all cases in RECS.)

Preliminary data release date: May 2022

Table HC6.1 Space heating in U.S. homes, by housing unit type, 2020

	Number of housing units (million)					
		Housing unit type				
	Total U.S.[a]	Single-family detached	Single-family attached	Apartments (2–4 unit building)	Apartments (5 or more unit building)	Mobile home
All homes	123.53	77.07	7.45	9.34	22.84	6.83
Space heating equipment						
Uses space heating equipment	117.43	74.67	6.98	8.70	20.47	6.60
Has space heating equipment but does not use it	4.13	1.54	0.35	0.44	1.70	Q
Does not have space heating equipment	1.97	0.86	0.11	0.21	0.66	0.13
Main heating fuel and equipment						
Natural gas	56.25	39.71	4.08	4.02	7.00	1.44
Central warm-air furnace	47.37	35.90	3.42	2.46	4.26	1.34
Steam or hot water system	6.37	2.60	0.48	1.19	2.04	Q
Built-in room heater	2.33	1.09	0.17	0.35	0.68	Q

For SAS users

SAS Example 1: Calculate the frequency and RSE of households that used natural gas as their main space heating fuel (Table HC6.1)

Step 1. Create a new variable to flag the records of 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 its main space heating fuel, and 0 otherwise.

```
DATA RECS20;
   SET RECS2020_PUBLIC_V1;
   IF FUELHEAT=1 THEN NG_MAINSPACEHEAT =1; ELSE
   NG_MAINSPACEHEAT =0;
RUN;
```

Step 2. 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 associated with the estimate, use PROC SURVEYFREQ to process the replicate weights.

```
PROC SURVEYFREQ DATA=RECS20 VARMETHOD=JK;
   REPWEIGHTS NWEIGHT1-NWEIGHT60;
   WEIGHT NWEIGHT;
   TABLES NG_MAINSPACEHEAT;
RUN;
```

Answer. The estimated number of households that used natural gas as their main space heating fuel is 56,245,388 households. The standard deviation of the frequency is 545,591, and the calculation for the RSE is (545,591 / 56,245,389)*100 = 0.97. In other words, the sampling error is about 1% of the estimate, a relatively small amount, indicating that the estimate is very precise.

Table of ng_mainspaceheat					
ng_mainspaceheat	Frequency	Weighted Frequency	Std Err of Wgt Freq	Percent	Std Err of Percent
0	9881	67283636	545591	54.4679	0.4417
1	8615	56245389	545591	45.5321	0.4417
Total	18496	123529025	0.14759	100.000	

For R users

First, install the *survey* and *dplyr* package (Lumley 2017):

```
install.packages("survey",”dplyr”)
library(survey)
library(dplyr)
```

Read in the CSV file:

```
RECS2020 <- read.csv(file='< location where file is stored >', header=TRUE, sep=",")
```

R Example 1: Calculate the frequency and RSE of households that used natural gas as their main space heating fuel (Table HC6.1)

Step 1. Create a new variable to flag the records of 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 its main space heating fuel, and 0 otherwise.

```
RECS2020$NG_MAINSPACEHEAT <- ifelse(RECS2020$FUELHEAT == 1, 1, 0)
```

Step 2. Define the Jackknife replicate weights used in estimation:

```
repweights<-select(RECS2020,NWEIGHT1:NWEIGHT60)
```

Step 3. Define the survey design with the Jackknife replicate weights to calculate appropriate standard errors:

```
RECS <- svrepdesign(data = RECS2020,
                    weight = ~NWEIGHT,
                    repweights = repweights,
                    type = "JK1",
                    combined.weights = TRUE,
                    scale = (ncol(repweights)-1)/ncol(repweights),
                    mse = TRUE)
```

Step 4. Use *svytotal* to sum the number of households by NG_MAINSPACEHEAT, using the survey design defined above.

```
NG_MAINSPACEHEAT<-as.data.frame(svytotal(~NG_MAINSPACEHEAT,RECS))
```

Answer. The estimated total of households that used natural gas as their main space heating fuel is 56,245,389 households. The calculation for the RSE is (545,591 / 56,245,389)*100 = 0.97. The sampling error is about 1% of the estimate, which is relatively small.

	total	SE
NG_MAINSPACEHEAT	56245389	545591

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). EIA recommends following these 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 on reported data (Z variable = 0) or was imputed (Z variable = 1). Variables from the RECS questionnaire that were not imputed, contained no missing data, or were not from the questionnaire have no corresponding Z variables. EIA recommends using the imputed data, where available, to avoid biased estimation.

3. *Standardized coding:* 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 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 2020. 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 2020 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 warm climate might have heating equipment but did not use it in 2020. For this case, ELWARM would be 0.

5. *Confidentiality:* The 2020 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. EIA took the following steps to avoid disclosure of personally identifiable information in the public use microdata file.

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

- The following variables were removed due to too few responses or for disclosure risk:
  - COMBINED (on-site combined heat and power)
  - WIND (on-site wind generation)
  - PVINSTALL (year PV was installed)
  - PVCAPACITY (capacity of PV system in kilowatts)
  - APTEVCHG (do respondents in apartment building with 5+ units have access to an electric vehicle [EV] charger)
  - EVMAKE, EVMODEL, EVYEAR (EV make, model, and year)
  - EVCHRGAPT, EVCHRGWKS, EVCHRGBUS, EVCHRGMUNI, EVCHRGHWY, EVHCRGOTH (respondent charged their EV at apartment building, work, a business or shopping center, a municipal parking lot, a highway rest stop, a car dealership, or somewhere else)
  - EVHOMEAMT (what percentage of EV charging was done at home)
  - EVCHRGTYPE (what type of EV charger does respondent have at home)
  - EVWRKMILES (average number of miles EV is driven a week)
- The following variables were top-coded:
  - BEDROOMS (number of bedrooms) was top-coded to 6.
  - OTHROOMS (number of other rooms) was top-coded to 9.
  - NCOMBATH (number of full bathrooms) was top-coded to 4.
  - NHAFBATH (number of half bathrooms) was top-coded to 2.
  - HHAGE (age of the householder) was top-coded to 90.
  - NHSLDMEM (number of household members) was top-coded to 7.
  - NUMCHILD (number of children under 18) was top-coded to 4.
- EIA added random errors to weather and climate (HDD30YR and CDD30YR) values. Adjustments were minor and will not result in significant differences for aggregate estimation.

References

Lohr, S.L. (2010). Sampling: Desing and Analysis. 2nd ed. Boston: Brooks/Cole. Page 380−383.

Lumley, T. (2017) "Survey: analysis of complex survey samples". R package version 4.1-1.

The R code presented in this document was developed and tested in version 4.2.0.