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

Published June 2022

Revised June 2023

Independent Statistics and Analysis
U.S. Energy Information Administration

www.eia.gov
U.S. Department of Energy
Washington, DC 20585

The U.S. Energy Information Administration (EIA), the statistical and analytical agency within the U.S. Department of Energy (DOE), prepared this report. By law, our data, analyses, and forecasts are independent of approval by any other officer or employee of the U.S. Government. The views in this report do not represent those of DOE or any other federal agencies.

Table of Contents

Overview

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

We 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, we 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, we 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 household represents.

As part of the weighting process, we first calculated base sampling weights, which are the reciprocal of the probability of being selected for the RECS sample, for each sampled housing unit. We then adjusted the base weights to account for survey nonresponse and eligibility. In addition, we used poststratification adjustments to ensure that the RECS weights add up to the estimated 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 responding household represents. For example, if NWEIGHT for a household is 10,000, that household represents itself and 9,999 other households in the population that either were not sampled or were sampled but did not respond to the survey. 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 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 include weighted estimates and RSEs. An RSE is formulated as the standard error (square root of the sampling variance) of a survey estimator, divided by the survey estimate, and multiplied by 100. In other words, the RSE quantifies how much the the estimator varies over all possible samples that could have been selected from the population using the same sample design, relative to the corresponding survey estimate, and expressed as a percentage. The smaller the RSE, the more precise a survey estimate is in terms of its sampling variability. An RSE for each estimate in the RECS tables is under a separate tab in the table. Estimates greater than zero but 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 shown below. Note that an RSE can be calculated by multiplying the coefficient of variation (CV) by 100 in the SAS/STAT statistical software.

Jackknife method of estimating standard error

The 2020 RECS uses the Jackknife method to produce replicate weights 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. We 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 parameter of interest, let $\widehat{\theta}$ be the estimate from the full sample for θ. Let $\hat{\theta}r$ be an estimator used for the r-th replicate, and R is the total number of the replicate weights, the variance of $\widehat{\theta}$ is estimated by:

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

The formula for calculating the RSE is:

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

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). Software packages such as SAS/STAT, R, Stata, SUDAAN, and WesVar can process replicate weights to calculate RSEs. We provided instructions for Excel users and users with access to SAS/STAT and R. Note that the version and components of SAS/STAT used could affect the analysis capability; examples used below were done in SAS/STAT 14.1. We show how to compute point estimates using Excel, but Excel does not have a built-in function that calculates RSEs directly using replicate weights. We recommend calculating standard errors or RSEs using the supplied replicate weights 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)

You can estimate a simple count of households 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), which results in 9,595 cases.

Step 2. Sum the NWEIGHT column for these 9,595 cases.

Answer: The estimated number of households that used natural gas as main heating fuel was approximately 62,713,449 households. This amount is equal to 51% of all homes, or 62.71 million/123.53 million (the sum of NWEIGHT for all cases in RECS.)

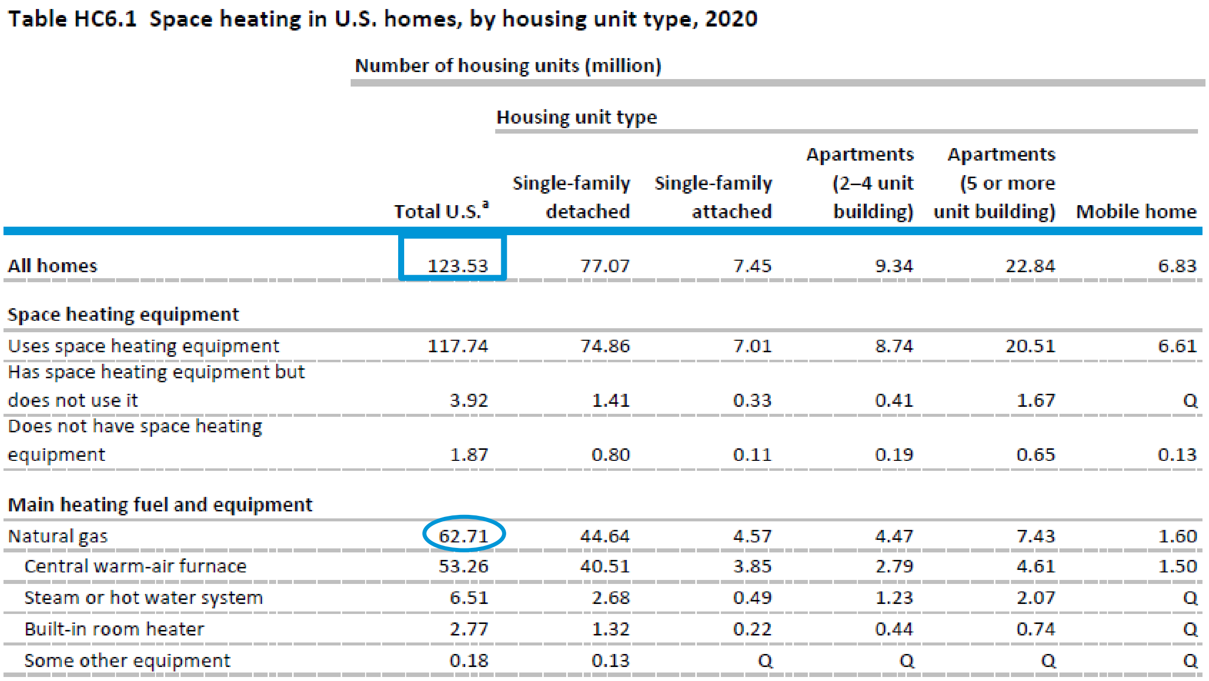


Table HC6.1 Space heating in U.S. homes, by housing unit type, 2020

	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	117.74	74.86	7.01	8.74	20.51	6.61
Has space heating equipment but does not use it	3.92	1.41	0.33	0.41	1.67	Q
Does not have space heating equipment	1.87	0.80	0.11	0.19	0.65	0.13
Main heating fuel and equipment						
Natural gas	62.71	44.64	4.57	4.47	7.43	1.60
Central warm-air furnace	53.26	40.51	3.85	2.79	4.61	1.50
Steam or hot water system	6.51	2.68	0.49	1.23	2.07	Q
Built-in room heater	2.77	1.32	0.22	0.44	0.74	Q
Some other equipment	0.18	0.13	Q	Q	Q	Q

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

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 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_NG;
   SET RECS2020_PUBLIC_V3;
   IF FUELHEAT=1 THEN NG_MAINSPACEHEAT=1;
   ELSE NG_MAINSPACEHEAT=0;
RUN;
```

Step 2.

Use the PROC SURVEYFREQ procedure with the VARMETHOD, WEIGHT, and REPWEIGHTS statements to obtain sampling errors associated with the estimates. The jackknife coefficient is 59/60, which is also the default value in the procedure; therefore, it does not need to be specified in the JKCOEFS option. In addition, for the population total estimate, the CLWT and CVWT options, respectively, provide the 95% confidence limits and the coefficient of variation (CV). Similarly, in obtaining the confidence limits and coefficient of variation for the percentages (proportions) associated with each category, use the CL and CV options.

```
PROC SURVEYFREQ DATA=RECS20_NG VARMETHOD=JK;
   REPWEIGHTS NWEIGHT1-NWEIGHT60;
   WEIGHT NWEIGHT;
   TABLES NG_MAINSPACEHEAT/CLWT CVWT CL CV;
RUN;
```

Answer. The estimated number of households that used natural gas as their main space-heating fuel is 62,713,449 households (with an estimated percentage of 50.8%). The RSE for the estimates is 0.0077 (CV) *100=0.77; or you can also calculate the RSE using the standard error of the frequency, which is 483,047, the RSE calculation of this approach is (483,047/62,713,449)*100 = 0.77. In other words, the relative standard error is less than 1% of the estimated total population, a relatively small amount, indicating that the estimate is very precise. Note that the estimates for NG_MAINSPACEHEAT = 0 reflect consumption for homes that do not use natural gas as the main space-heating fuel.

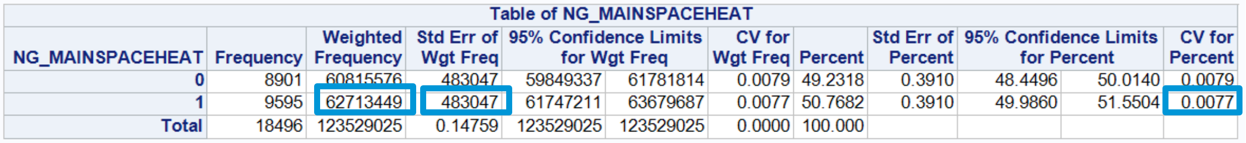


Table of NG_MAINSPACEHEAT

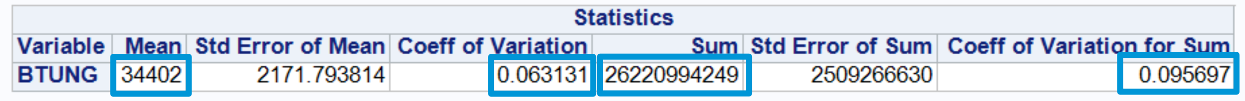
NG_MAINSPACEHEAT	Frequency	Weighted Frequency	Std Err of Wgt Freq	95% Confidence Limits for Wgt Freq		CV for Wgt Freq	Percent	Std Err of Percent	95% Confidence Limits for Percent		CV for Percent
0	8901	60815576	483047	59849337	61781814	0.0079	49.2318	0.3910	48.4496	50.0140	0.0079
1	9595	62713449	483047	61747211	63679687	0.0077	50.7682	0.3910	49.9860	51.5504	0.0077
Total	18496	123529025	0.14759	123529025	123529025	0.0000	100.000				

SAS Example 2: Calculate the sum and average of the total natural gas consumption used for the households in South Carolina (SC) (Table CE4.1.NG.ST *Annual household site natural gas consumption in the United States by end use – totals and percentages, 2020*).

To calculate the sum and average of total natural gas consumption, first filter the dataset by BTUNG>0. The VARMETHOD, WEIGHT, and REPWEIGHT statements are the same as the PROC SURVEYFREQ example above. The SUM and MEAN options provide the estimates of the sum and the average. Use the WHERE statement to specify state_postal='*SC*' in this example.

```
DATA RECS2020_NGTOTALUSED;
   SET RECS2020_PUBLIC_V3;
   IF BTUNG>0;
RUN;

PROC SURVEYMEANS DATA=RECS2020_NGTOTALUSED VARMETHOD=JK MEAN SUM CV
CVSUM;
   REPWEIGHTS NWEIGHT1-NWEIGHT60;
   WEIGHT NWEIGHT;
   VAR BTUNG;
   WHERE state_postal='SC';
RUN;
```

Answer. The estimated total consumptions of households that used natural gas in SC is 26.2 trillion British thermal units (Btu), with RSE=9.6 (CV*100). The average natural gas consumption used is 34.4 million Btu, with RSE=6.3. The table of output below shows the results.



Statistics						
Variable	Mean	Std Error of Mean	Coeff of Variation	Sum	Std Error of Sum	Coeff of Variation for Sum
BTUNG	34402	2171.793814	0.063131	26220994249	2509266630	0.095697

SAS Example 3: Calculate the energy intensity per square foot by climate zone for the United States (Table CE1.1)

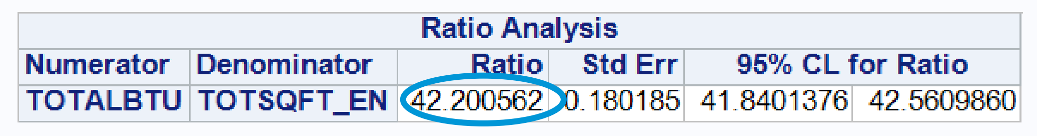
Step 1. Create a new variable called Climate_region to combine climate zones.

```
DATA RECS20_NG_CLIMATE;
   SET RECS20_NG;
   length Climate_Region $20.;
   If BA_climate in ("Subarctic", "Very-Cold", "Cold") then
   Climate_Region="Very cold/Cold";
   if BA_climate in ("Mixed-Humid") then Climate_Region="Mixed-
   humid";
   if BA_climate in ("Mixed-Dry", "Hot-Dry") then
   Climate_Region="Mixed-dry/Hot-dry";
   if BA_climate in ("Hot-Humid") then Climate_Region="Hot-humid";
   if BA_climate in ("Marine") then Climate_Region="Marine";
RUN;
```

Step 2. To calculate the energy intensity in SAS, use the SURVEYMEANS procedure and the RATIO statement. For this example, use BA_climate in the DOMAIN statement, the TOTALBTU and TOTSQFT_EN in the RATIO statement to calculate the intensity per square foot. The WEIGHT and REPWEIGHT variables are the same as the examples above. Use the ODS SELECT statement to select the desired output for display.

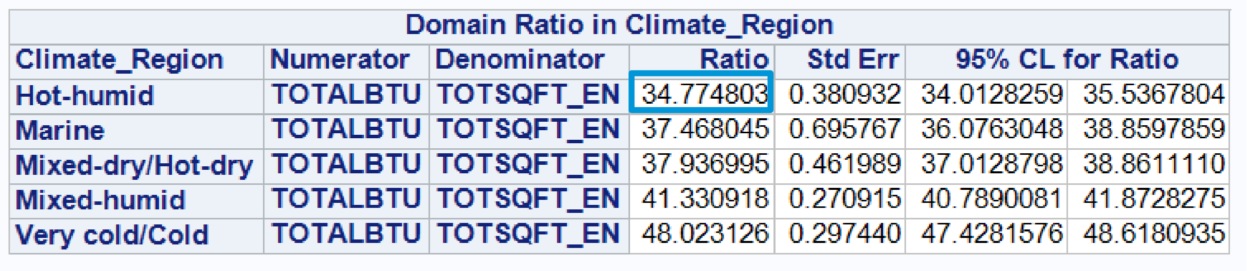
```
PROC SURVEYMEANS DATA=RECS20_NG_CLIMATE VARMETHOD=JK MEAN CLM;
   REPWEIGHTS NWEIGHT1-NWEIGHT60;
   WEIGHT NWEIGHT;
   DOMAIN Climate_Region;
   RATIO TOTALBTU/TOTSQFT_EN;
   ODS SELECT RATIO DOMAINRATIO;
RUN;
```

The first Ratio Analysis table shows the intensity for all U.S. homes, which is about 42.2 trillion Btu, this is, U.S. homes use about 42,000 British thermal units (Btu) per square foot. The intensity by climate zones can be found in the *Domain Ratio in Climate_Region* table. For example, the intensity per square foot in the hot-humid region is about 35,000 Btu.



Ratio Analysis

Numerator	Denominator	Ratio	Std Err	95% CL for Ratio	
TOTALBTU	TOTSQFT_EN	42.200562	0.180185	41.8401376	42.5609860



Domain Ratio in Climate_Region

Climate_Region	Numerator	Denominator	Ratio	Std Err	95% CL for Ratio	
Hot-humid	TOTALBTU	TOTSQFT_EN	34.774803	0.380932	34.0128259	35.5367804
Marine	TOTALBTU	TOTSQFT_EN	37.468045	0.695767	36.0763048	38.8597859
Mixed-dry/Hot-dry	TOTALBTU	TOTSQFT_EN	37.936995	0.461989	37.0128798	38.8611110
Mixed-humid	TOTALBTU	TOTSQFT_EN	41.330918	0.270915	40.7890081	41.8728275
Very cold/Cold	TOTALBTU	TOTSQFT_EN	48.023126	0.297440	47.4281576	48.6180935

SAS Example 4: Compare if the proportions and the average energy consumption of households using natural gas as their main space-heating fuel are statistically different among the households in the Census regions.

To compare if the proportions of households using natural gas as their main space-heating fuel are different among the Census regions, use the CHISQ or WCHISQ options in the Tables statement of the PROC SURVEYFREQ procedure to test the association between the two variables.

```
PROC SURVEYFREQ DATA=RECS20_NG VARMETHOD=JK;
   REPWEIGHTS NWEIGHT1-NWEIGHT60;
   WEIGHT NWEIGHT;
   TABLES NG_MAINSPACEHEAT*REGIONC/CHISQ WCHISQ;
 RUN;
```

The CHISQ option provides the Rao-Scott adjusted Chi-square statistics for testing associations between the use of natural gas main space heating and region. The WCHISQ option computes the Wald Chi-square test. In this example, the P value from each of the test is <.0001, therefore, indicating the use of natural gas main space heating is dependent of the Region variable.

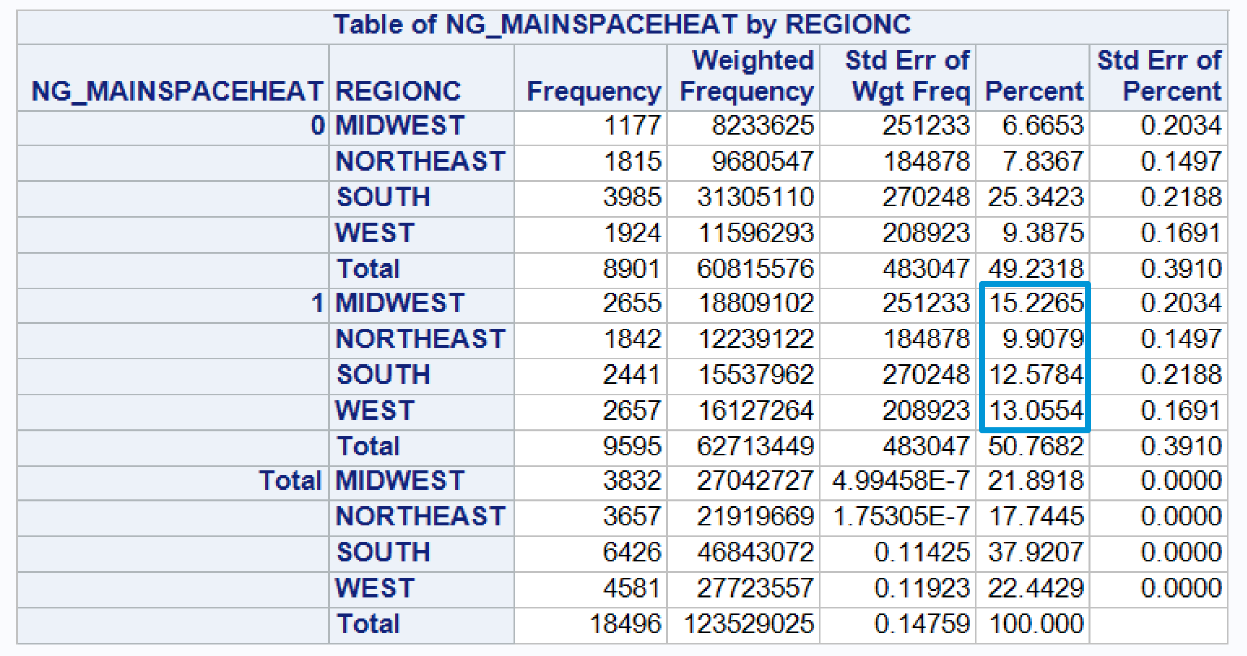
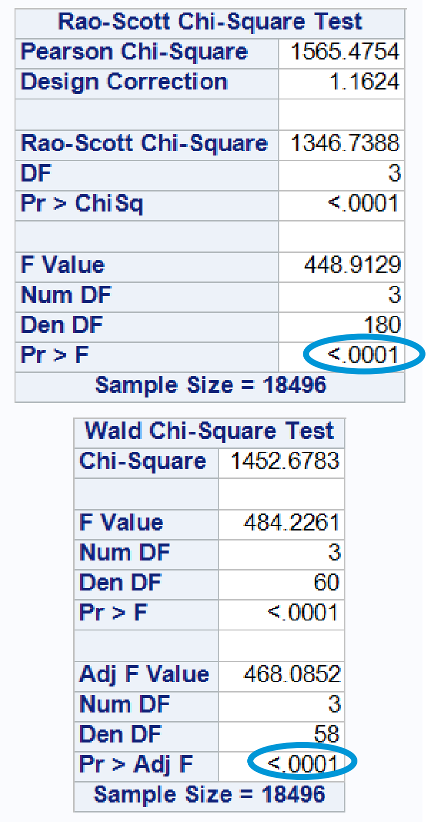


Table of NG_MAINSPACEHEAT by REGIONC

NG_MAINSPACEHEAT	REGIONC	Frequency	Weighted Frequency	Std Err of Wgt Freq	Percent	Std Err of Percent
0	MIDWEST	1177	8233625	251233	6.6653	0.2034
	NORTHEAST	1815	9680547	184878	7.8367	0.1497
	SOUTH	3985	31305110	270248	25.3423	0.2188
	WEST	1924	11596293	208923	9.3875	0.1691
	Total	8901	60815576	483047	49.2318	0.3910
1	MIDWEST	2655	18809102	251233	15.2265	0.2034
	NORTHEAST	1842	12239122	184878	9.9079	0.1497
	SOUTH	2441	15537962	270248	12.5784	0.2188
	WEST	2657	16127264	208923	13.0554	0.1691
	Total	9595	62713449	483047	50.7682	0.3910
Total	MIDWEST	3832	27042727	4.99458E-7	21.8918	0.0000
	NORTHEAST	3657	21919669	1.75305E-7	17.7445	0.0000
	SOUTH	6426	46843072	0.11425	37.9207	0.0000
	WEST	4581	27723557	0.11923	22.4429	0.0000
	Total	18496	123529025	0.14759	100.000	



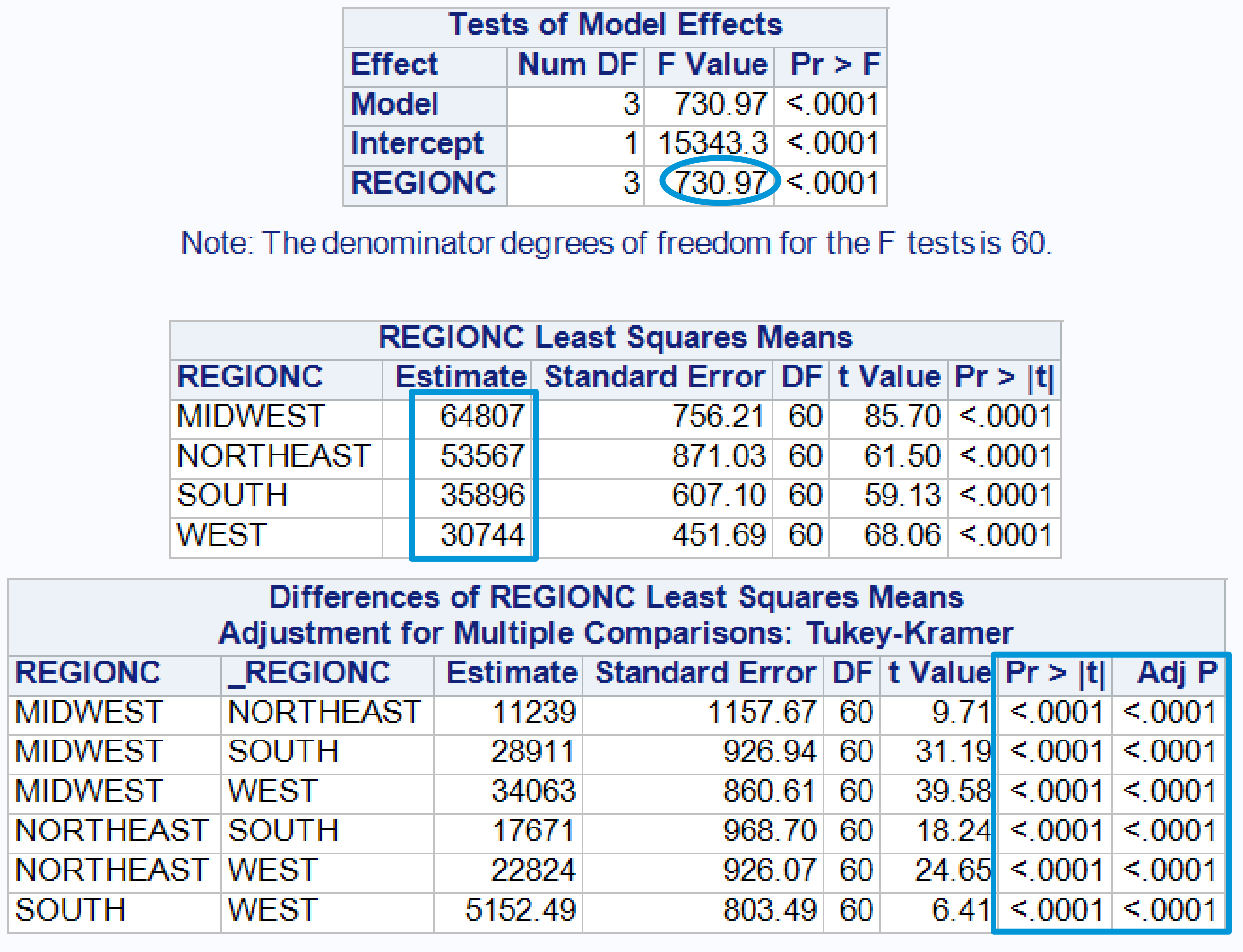
Rao-Scott Chi-Square Test	
Pearson Chi-Square	1565.4754
Design Correction	1.1624
Rao-Scott Chi-Square	1346.7388
DF	3
Pr > ChiSq	<.0001
F Value	448.9129
Num DF	3
Den DF	180
Pr > F	<.0001
Sample Size = 18496	

Wald Chi-Square Test	
Chi-Square	1452.6783
F Value	484.2261
Num DF	3
Den DF	60
Pr > F	<.0001
Adj F Value	468.0852
Num DF	3
Den DF	58
Pr > Adj F	<.0001
Sample Size = 18496	

To compare if the average space-heating consumption for households using natural gas as their main space-heating fuel are different among the Census regions, use the DOMAIN, CLASS, and LSMEANS statement in the PROC SURVEYREG statement.

```
PROC SURVEYREG DATA=RECS20_NG VARMETHOD=JK;
   REPWEIGHTS NWEIGHT1-NWEIGHT60;
   WEIGHT NWEIGHT;
   WHERE NG_MAINSPACEHEAT=1;
   CLASS REGIONC;
   MODEL TOTALBTUSPH=REGIONC;
   LSMEANS REGIONC/ADJUST=TUKEY;
RUN;
```

The LSMEANS statement in the SURVEYREG procedure provides the least square means for each region, and the ADUST=TUKEY option in the LSMEANS statement provides the results of the F-Statistics of the model and several comparisons by the TUKEY method. The results in the *Adj P* column of the bottom table below indicate that all pairs of Census regions are statistically different in average space-heating consumption.



Tests of Model Effects

Effect	Num DF	F Value	Pr > F
Model	3	730.97	<.0001
Intercept	1	15343.3	<.0001
REGIONC	3	730.97	<.0001

Note: The denominator degrees of freedom for the F tests is 60.

REGIONC Least Squares Means

REGIONC	Estimate	Standard Error	DF	t Value	Pr > |t|
MIDWEST	64807	756.21	60	85.70	<.0001
NORTHEAST	53567	871.03	60	61.50	<.0001
SOUTH	35896	607.10	60	59.13	<.0001
WEST	30744	451.69	60	68.06	<.0001

Differences of REGIONC Least Squares Means
Adjustment for Multiple Comparisons: Tukey-Kramer

REGIONC	_REGIONC	Estimate	Standard Error	DF	t Value	Pr > |t|	Adj P
MIDWEST	NORTHEAST	11239	1157.67	60	9.71	<.0001	<.0001
MIDWEST	SOUTH	28911	926.94	60	31.19	<.0001	<.0001
MIDWEST	WEST	34063	860.61	60	39.58	<.0001	<.0001
NORTHEAST	SOUTH	17671	968.70	60	18.24	<.0001	<.0001
NORTHEAST	WEST	22824	926.07	60	24.65	<.0001	<.0001
SOUTH	WEST	5152.49	803.49	60	6.41	<.0001	<.0001

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 (note that using the sas7bdat data file might produce slight differences in some results due to variable formatting):

```
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 you will use for estimation:

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

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

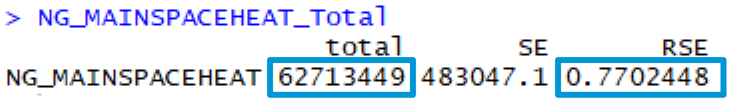
```
RECS <- svrepdesign(data = RECS2020,
                    weight = ~NWEIGHT,
                    repweights = repweights,
                    type = "JK1",
                    combined.weights = TRUE,
                    scale = (ncol(repweights)-1)/ncol(repweights),
                    mse = TRUE)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

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

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

Answer. The estimated total households that used natural gas as their main space-heating fuel is 62,713,449 households. The calculation for the RSE is (483,047 / 62,713,449)*100 = 0.77. The sampling error is less than 1% of the estimate, which is relatively small. Alternatively, the RSE can be derived from:

```
NG_MAINSPACEHEAT_Total$RSE<-
(NG_MAINSPACEHEAT_Total$SE/NG_MAINSPACEHEAT_Total$total)*100
```



```
> NG_MAINSPACEHEAT_Total
                    total       SE       RSE
NG_MAINSPACEHEAT 62713449 483047.1 0.7702448
```

To obtain the proportion estimate, use the *svymean()* function instead of the *svytotal* in the expression in Step 4. In addition, the *confint()* function provides the 95% confidence limits.

R Example 2: Calculate the sum and average of the total natural gas used for the households in South Carolina (SC) (Table CE4.1.NG.ST *Annual household site natural gas consumption in the United States by end use – totals and percentages, 2020*).

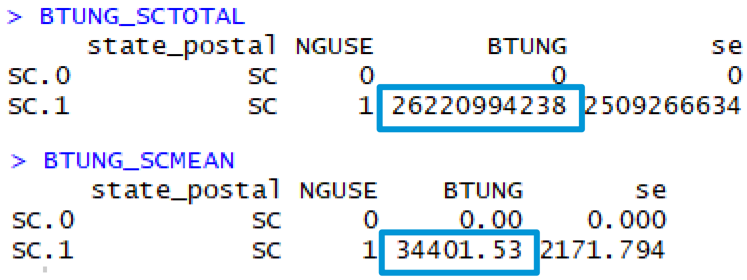
To calculate the total consumption estimates in R, use the *svytotal()* function; and use the *svymean()* function for the average consumption . In addition, use *svyby()* to group households by USENG and state (state_postal) and the *subset()* function to limit the results to SC only.

First, create a new variable to flag the households that have positive natural gas consumption for any natural gas end use. This new variable NGUSE is equal to 1 if BTUNG is greater than 0 and 0 otherwise. Then, run the survey design for the dataset again before producing estimates using the functions mentioned above.

```
RECS2020$NGUSE <- ifelse(RECS2020$BTUNG > 0, 1,0)

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

```
# calculate the total:

BTUNG_TOTAL<-svyby(~BTUNG, by=~state_postal+NGUSE, RECS, svytotal)

BTUNG_SCTOTAL<-subset(BTUNG_TOTAL, state_postal=='SC')


# calculate the mean:

BTUNG_MEAN<-svyby(~BTUNG, by=~state_postal+NGUSE, RECS, svymean)

BTUNG_SCMEAN<-subset(BTUNG_MEAN, state_postal==’SC’)
```

The output below shows the result for SC. The total estimated consumption for households that used natural gas in SC is 26.2 trillion British thermal units (Btu). The RSE for the total is (2509266634/26220994238)*100 = 9.6%. The average consumption per household is 34.4 million BTU, with RSE=6.3. As mentioned in R example 1, the 95% confidence limits with the *conflint()* function. Note that the estimates for NGUSE = 0 reflect consumption for homes that do not use any natural gas.



```
> BTUNG_SCTOTAL
     state_postal NGUSE         BTUNG          se
SC.0           SC     0             0           0
SC.1           SC     1 26220994238 2509266634

> BTUNG_SCMEAN
     state_postal NGUSE    BTUNG       se
SC.0           SC     0     0.00    0.000
SC.1           SC     1 34401.53 2171.794
```

R Example 3: Calculate the energy intensity per square foot by climate zone for the United States (Table CE1.1)

Step 1. Create a new variable called Climate_region to combine climate zones, and rerun the survey design RECS.

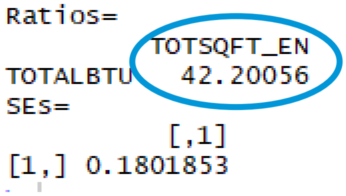
```
RECS2020$Climate_Region <- as.factor(ifelse(RECS2020$BA_climate=='Subarctic', 'Very cold/Cold',
                          ifelse(RECS2020$BA_climate=='Very-Cold', 'Very cold/Cold',
                          ifelse(RECS2020$BA_climate=='Cold', 'Very cold/Cold',
                          ifelse(RECS2020$BA_climate=='Mixed-Humid', 'Mixed-humid',
                          ifelse(RECS2020$BA_climate=='Mixed-Dry', 'Mixed-dry/Hot-dry',
                          ifelse(RECS2020$BA_climate=='Hot-Dry', 'Mixed-dry/Hot-dry',
                          ifelse(RECS2020$BA_climate=='Hot-Humid', 'Hot-humid',
                          ifelse(RECS2020$BA_climate=='Marine', 'Marine', NA))))))))
```

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

Step 2. To calculate the energy intensity per square foot for all U.S. homes, use the *svyratio()* function.

```
BTUPERSQFT<-svyratio(~TOTALBTU, ~TOTSQFT_EN, RECS)
```

The national estimate for energy intensity per square foot is about 42,000 Btu, as shown in Table CE1.1. The RSE is (0.1801853/42.20056)*100 = 0.43.

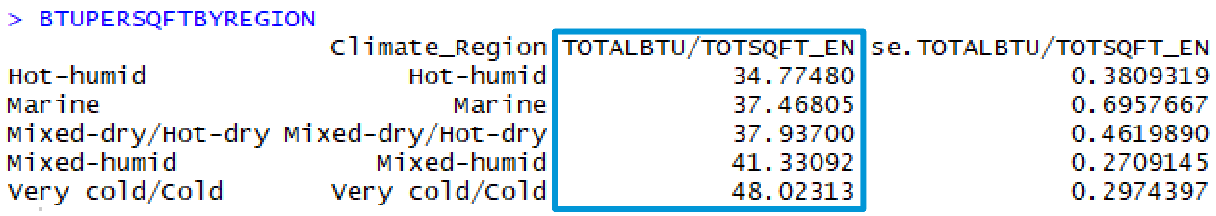


```
Ratios=
          TOTSQFT_EN
TOTALBTU    42.20056
SEs=
          [,1]
[1,] 0.1801853
```

To calculate the regional energy intensity per square foot, use svyratio with *svyby()*.

BTUPERSQFTBYREGION<-svyby(~TOTALBTU, by=~Climate_Region,denominator=~TOTSQFT_EN, RECS, svyratio)

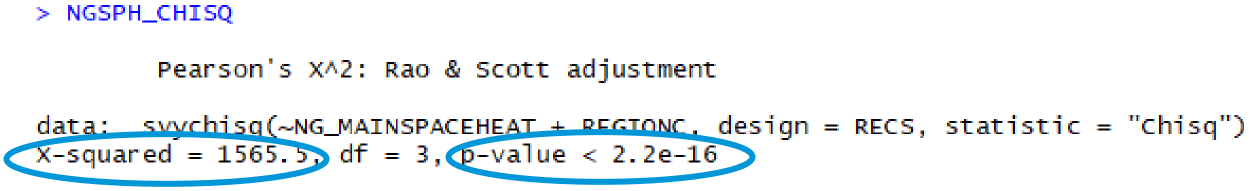
As an example, the average total consumption per square foot in the hot-humid climate is about 35,000 Btu, as shown in the table below.



```
> BTUPERSQFTBYREGION
                        Climate_Region TOTALBTU/TOTSQFT_EN se.TOTALBTU/TOTSQFT_EN
Hot-humid                    Hot-humid            34.77480              0.3809319
Marine                          Marine            37.46805              0.6957667
Mixed-dry/Hot-dry  Mixed-dry/Hot-dry            37.93700              0.4619890
Mixed-humid                Mixed-humid            41.33092              0.2709145
Very cold/Cold          Very cold/Cold            48.02313              0.2974397
```

R Example 4: Compare if the proportions and the consumption means for households using natural gas as their main space-heating fuel are statistically different among the households in different Census regions.

To compare if the proportions of households using natural gas as their main space-heating fuel are different among the Census regions, use the *svychisq()* function to obtain chi-square statistics. *NGSPH_CHISQ<-svychisq(~NG_MAINSPACEHEAT+REGIONC,design=RECS,statistic="Chisq")*

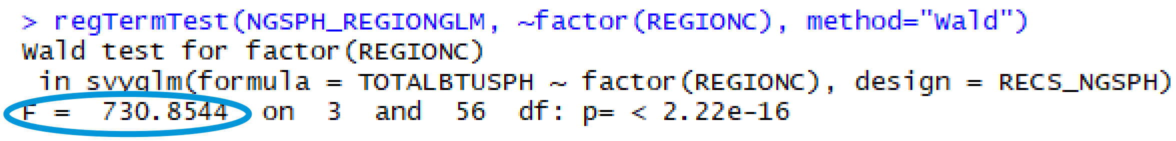


```
> NGSPH_CHISQ

        Pearson's X^2: Rao & Scott adjustment

data:  svychisq(~NG_MAINSPACEHEAT + REGIONC, design = RECS, statistic = "Chisq")
X-squared = 1565.5, df = 3, p-value < 2.2e-16
```

To compare if the average space-heating consumption estimates for households using natural gas as their main space-heating fuel are different among the Census regions, use the *svyglm()* function to run a regression model and obtain the coefficient, and use the *regTermTest()* function to obtain the F statistics.

RECS_NGSPH<-subset(RECS,NG_MAINSPACEHEAT==1)

NGSPH_REGIONGLM<-svyglm(TOTALBTUSPH~factor(REGIONC), design=RECS_NGSPH)
regTermTest(NGSPH_REGIONGLM, ~factor(REGIONC), method="Wald")



```
> regTermTest(NGSPH_REGIONGLM, ~factor(REGIONC), method="Wald")
Wald test for factor(REGIONC)
 in svyglm(formula = TOTALBTUSPH ~ factor(REGIONC), design = RECS_NGSPH)
F =  730.8544  on  3  and  56  df: p= < 2.22e-16
```

Notes to Consider When Using the Microdata File and Replicate Weights

1. *Publication standards:* We do not publish RECS estimates where the RSE is higher than 50 or the number of households used for the calculation is less than 10 (indicated by a *Q* in the data tables). We recommend following these guidelines for custom analysis using the public use microdata file.

2. *Imputation variables:* We imputed most variables for *Don't Know* and *Refuse* responses. The *Z variables*, also referred to as *imputation flags*, are in the public use microdata file. The imputation flag indicates whether we based the corresponding non-Z variable was reported data (Z variable = 0) or if we imputed it (Z variable = 1). Variables from the RECS questionnaire that we did not impute, contained no missing data, or were not from the questionnaire have no corresponding *Z* variables. We recommend using the imputed data, where available, to avoid biased estimation.

3. *Standardized coding:* Variables that we did not ask all respondents use the response code –2 for *Not Applicable*. For example, respondents who answered that they did not use any televisions at home (TVCOLOR = 0) were not asked what size television they most use at home, so TVSIZE1 = -2. Use caution when performing calculations on variables that 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. We derived these variables from answers given by each respondent, and they indicate whether the respondent had access to the fuel, used the fuel, and engaged in a specific end use. All indicators are either a 0 or a 1 for each combination of major fuel and end use. For example, respondents who say they heated their homes with electricity in 2020 will have the derived variable ELWARM = 1. If respondents say they have equipment but did not use it, the corresponding indicator is 0. As an example, respondents in a warm climate might have heating equipment but did not use it in 2020. For this case, ELWARM is 0.

5. *Confidentiality:* We collected the 2020 RECS under the authority of the Confidential Information Protection and Statistical Efficiency Act (CIPSEA). The agency, project staff, and our contractors and agents are personally accountable for protecting the identity of individual respondents. We took the following steps to avoid disclosing personally identifiable information in the public-use microdata file.

- We removed local geographic identifiers of sampled housing units, such as addresses.
- We removed the following variables because we received too few responses or because we found a disclosure risk:
  - COMBINED (on-site combined heat and power)
  - WIND (on-site wind generation)
  - PVINSTALL (year photovoltaic solar [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 an EV at their apartment building, place of 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) to 6
  - OTHROOMS (number of other rooms) to 9
  - NCOMBATH (number of full bathrooms) to 4
  - NHAFBATH (number of half bathrooms) to 2
  - HHAGE (age of the householder) to 90
  - NHSLDMEM (number of household members) to 7
  - NUMCHILD (number of children under 18) to 4
- We added random errors to weather and climate (HDD30YR and CDD30YR) values, as well as to the annualized consumption variables for electricity and natural gas.

Adjustments were minor and do not result in significant differences for aggregate estimation.

- We rounded the SQFTEST, TOTSQFT_EN, TOTHSQFT, and TOTCSQFT variables to the nearest 10.

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 SAS code and output for this paper was generated using SAS Enterprise Guide, and 14.1 Version of the SAS/STAT software. Copyright © 2017 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.

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