4. PREPARING ACS PUMS DATA FILES FOR ANALYSIS

The data dictionary available on the PUMS Technical Documentation Web page includes the complete list of variables in the American Community Survey (ACS) Public Use Microdata Sample (PUMS) data sets and additional documentation to help data users work with the files (see Figure 4.1).[20]

[20] U.S. Census Bureau, American Community Survey (ACS), PUMS Technical Documentation, <www.census.gov/programs-surveys/acs/technical-documentation/pums/documentation.html>.

Figure 4.1. **ACS PUMS Technical Documentation**

U.S. Census Bureau, American Community Survey (ACS), PUMS Technical Documentation, <www.census.gov/programs-surveys/acs/technical-documentation/pums/documentation.html>.

The PUMS Technical Documentation is organized by year of data release and includes several useful resources:

- PUMS ReadMe is updated with each data release and includes information about geography and variable changes, as well as guidance for getting started in using the ACS PUMS files.
- Subjects in the PUMS provides a list of the subjects that are included in the ACS PUMS files.
- PUMS Data Dictionary includes a list of all the variables in the ACS PUMS files, their values, and descriptions of what those values mean. The ACS PUMS data set includes variables for nearly every topic included in the ACS, as well as variables that were created by combining multiple survey responses.
- Code Lists provides detailed codes for variables with a long list of coded responses (for example, occupation or ancestry).
- PUMS Top Coded and Bottom Coded Values is a list of all the variables that have been top- and/or bottom-coded, and a list of state-specific values for the top- and bottom-coded variables. (More information about top-coding and bottom-coding is provided below.)
- Accuracy of the PUMS explains the sample design, estimation methodology, and the accuracy of the data. This section also includes instructions for calculating standard errors and margins of error for ACS PUMS estimates. (See the section on "Data Quality in the ACS PUMS" for more information.)
- PUMS Estimates for User Verification are weighted national- and state-level estimates, standard errors, and margins of error for several characteristics that data users can use to confirm that PUMS data files have been set up correctly.

PUMS data can be analyzed using a variety of statistical software packages (for example, SPSS, SAS, R, or Stata). In the examples below, SAS programming code is used to prepare the ACS PUMS data for analysis.

Working with Data for the Entire United States

The national ACS PUMS data sets contain millions of records and are very large. As a result, data sets for the entire United States are split into multiple files—"a" and "b" files for the ACS 1-year PUMS, and "a," "b," "c," and "d" files for the ACS 5-year PUMS. Data users must combine these component data files together to create a complete data set for the United States.

Here is a SAS statement that concatenates the "a" and "b" files for the ACS 1-year PUMS population records:

```
/*Concatenate a and b population records to obtain
all U.S. PUMS person records*/
data us_pums_person_data;
set psam_pusa psam_pusb;
run;
```

Data for individual states, the District of Columbia, and for Puerto Rico are also available as separate population data sets. When working with data for a single state, there are no "a" and "b" files, so data users can skip this step.

Combining Housing and Population Data

While some analyses can be conducted using only the housing data or population data, there are other instances where data users need to link housing unit records to population records. For example, housing unit records include data on the number of vehicles available, but to identify the characteristics of people who may have access to those vehicles, you need to link the housing unit records to the population records.

To combine the housing unit and population records, merge the two data sets together by matching on the SERIALNO variable. The SERIALNO variable is unique for each housing unit record across the nation. Matching on this variable merges the housing unit variables onto the population records.

Here is a portion of a SAS program that merges housing and population records:

```
/*Concatenate a and b housing records to obtain all
US PUMS housing records*/
data us_pums_housing_data;
 set psam_husa psam_husb;
run;

/*Sort the housing and population records by
SERIALNO*/
proc sort data=us_pums_housing_data;
by SERIALNO;
run;
proc sort data=us_pums_person_data;
by SERIALNO;
run;

/*merge the housing and population records
together*/
data merged;
merge us_pums_person_data us_pums_housing_data;
by SERIALNO;
run;
```

Merging the housing and population records yields a population-level data set that can be used to estimate the number or share of people with various household-based characteristics.[21]

Within households, data for each household member are included in separate person-level records. However, data users may be interested in producing estimates that combine data from multiple household members. For example, you may want to know the proportion of spouses who have the same level of educational attainment. The steps involved in joining these records are shown in Appendix A.

[21] Note that the housing unit data contain vacant housing units. These records will not have any corresponding person records.

Selecting Appropriate Weights

Each housing and person record is assigned a weight, because the records in the PUMS files represent a sample of the population. The weight is a numeric variable expressing the number of housing units or people that an individual microdata record represents. The sum of the housing unit and person weights for a geographic area is equal to the estimate of the total number of housing units and people in that area. To generate estimates based on the PUMS records, data users must correctly apply weights.

TIP: To generate statistics for housing units or households (for example, data on average household income), data users should apply the PUMS household weights (WGTP). To generate statistics for individuals (such as age or educational attainment), data users should apply the PUMS person weights (PWGTP).

When working with a merged file that includes both housing and person records, person weights should be used to produce estimates for person characteristics. Housing characteristics cannot be tallied from this merged file without taking extra steps to ensure that each housing weight is counted only once per household.

There are two additional sets of weights, one for households ranging from WGTP1 to WGTP80, and one for individuals ranging from PWGTP1 to PWGTP80. These "replicate weights" are used to calculate the error associated with each estimate. For more information about replicate weights, see the section below on "Data Quality in the ACS PUMS."

Selecting Variables

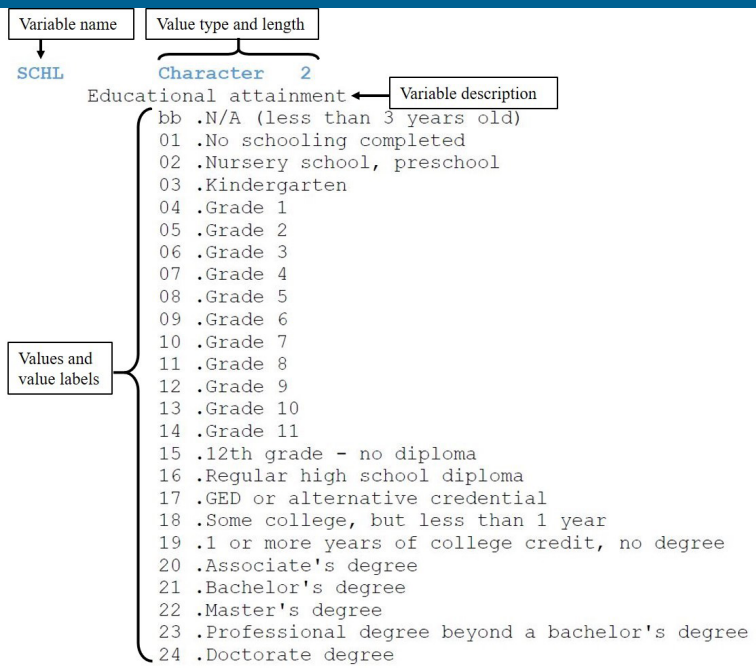
The PUMS Data Dictionary is available for each ACS PUMS data set and includes a complete list of the variables in the PUMS data files.[22] For each variable, the data dictionary includes the variable name, value type and length, variable description, values, and value labels (see Figure 4.2). The "bb" values indicate that data on educational attainment are not available for persons under the age of 3.

Some variables, such as ancestry, have a large number of coded responses. In addition to the data dictionary, values and value labels for these variables are maintained in a separate Excel file listed under "Code Lists" on the PUMS Technical Documentation Web page.[23] The Code Lists show the detailed ACS codes that are included in each value for a PUMS variable. For example, the "German" ancestry code in the ACS PUMS includes a range of related responses, such as "German," "Saxon," and "West German" (see Table 4.3).

[22] U.S. Census Bureau, American Community Survey (ACS), PUMS Technical Documentation, <www.census.gov/programs-surveys/acs/technical-documentation/pums/documentation.html>.

[23] U.S. Census Bureau, American Community Survey (ACS), PUMS Technical Documentation, <www.census.gov/programs-surveys/acs/technical-documentation/pums/documentation.html>.

Figure 4.2. **Elements of a Sample Variable in the ACS PUMS Data Dictionary**



Source: U.S. Census Bureau, 2017 ACS PUMS Data Dictionary, <www.census.gov/programs-surveys/acs/technical-documentation/pums/documentation.2017.html>.

Table 4.3. **German Ancestry Codes in the 2017 ACS 1-Year PUMS**

PUMS code	Ancestry description	Ancestry code	Corresponding detailed ancestry code
032	German	032	German
		033	Bavaria
		034	Berlin
		035	Hamburg
		036	Hannover
		037	Hessian
		038	Lubecker
		039	Pomeranian
		041	Saxon
		042	Sudetenlander
		043	Westphalian
		044	East German
		045	West German

Source: U.S. Census Bureau, American Community Survey, PUMS Technical Documentation, 2017 ACS 1-year PUMS Code Lists, <www.census.gov/programs-surveys/acs/technical-documentation/pums/documentation.2017.html>.

While the PUMS data dictionary provides the coded responses for each variable, the ACS Subject Definitions provide information about the meaning of each ACS variable or subject.[24] The subject definitions include information about how variables are defined and measured, the population universe, the survey questions used to derive each variable, how variables may have changed over time, and the comparability with ACS and decennial census variables from previous years.[25] These definitions apply to both ACS PUMS data, as well as pretabulated ACS data on the U.S. Census Bureau's Web site.

Top-Coded and Bottom-Coded Variables

ACS responses are strictly confidential. In addition to removing all identifying information, responses to open-ended questions, such as age, income, and housing unit value—where an extreme value might identify an individual—are top-coded and/or bottom-coded. Top-coding is the process of taking any response exceeding a particular value and replacing it with a predetermined value. These predetermined values differ by state. For example, for 2017, if someone in New York reports their age as 103, it will be reported in the ACS PUMS file as 95 (the top-coded value shown for New York). Refer to PUMS Top Coded and Bottom Coded Values in the PUMS Technical Documentation for the list of impacted variables and the predetermined values in each state.[26]

[24] U.S. Census Bureau, American Community Survey (ACS), Code Lists, Definitions, and Accuracy, <www.census.gov/programs-surveys/acs/technical-documentation/code-lists.html>.

[25] The Census Bureau's subject definitions were created primarily for use with published tables, and some of these definitions are not applicable to the variables in the PUMS.

[26] U.S. Census Bureau, American Community Survey (ACS), PUMS Technical Documentation, <www.census.gov/programs-surveys/acs/technical-documentation/pums/documentation.html>.

PUMS Estimates for User Verification

PUMS data users are responsible for the accuracy of their estimates. Because PUMS data consist of a subset of the full ACS sample, tabulations from the ACS PUMS will not match those from published tables of ACS data. You can verify that you have correctly accessed and tabulated data from the ACS PUMS file by replicating the values presented in "PUMS Estimates for User Verification" in the PUMS Technical Documentation.[27]

The PUMS estimates for user verification include weighted estimates for selected characteristics and associated standard errors and margins of error. The standard errors and margins of errors were calculated using the Successive Difference Replicate method (described in the next section). The estimates are produced for the United States and for each state, the District of Columbia, and for Puerto Rico. They are available in SAS and comma-separated value (CSV) formats.

Beginning in 2017, the user verification files for the ACS PUMS also include a CSV file with unweighted counts of the number of records in each PUMS file. Data users may verify that the number of records in their PUMS person or housing file match the number given in this file.[28]

[27] U.S. Census Bureau, American Community Survey (ACS), PUMS Technical Documentation, <www.census.gov/programs-surveys/acs/technical-documentation/pums/documentation.html>.

[28] Previously, the number of records for the United States and Puerto Rico combined was provided at the beginning of the PUMS "Accuracy of the PUMS" document.