The Weighting is the Hardest Part

Gregg Bell, Ph.D.

gbell@cba.ua.edu

CULVERHOUSE

COLLEGE OF COMMERCE

THE UNIVERSITY OF ALABAMA®

The Weighting is the Hardest Part:

An example of the correct use of P.U.M.S. data files

(Even for people who have never *really* used SAS!)

2014 ACS Data Users Conference.

May 29-30, 2014
Holiday Inn Capitol, Washington DC

Abstract:

The P.U.M.S. dataset is a rich source of valuable information, but using the data is not as straightforward as with many other Census Bureau data products because of the weights involved. A working knowledge of weights and weighting is essential to the correct analysis of these data. This short course will explain a little of the "why" and a lot of the "how" with regards to weights and their use using popular commercial software.

Introduction:

A simple example of cell weights:

Weights are generated to compensate for different response rates from different groups when it is known that members of the different groups respond similarly to other members of that group, but respond differently from the members of other groups. These weights are commonly referred to as "cell weights."

If gender, for instance, is considered a potential source of bias, the researcher may use gender as a weight generating factor.

For example, if we know that the male/female breakdown for the population is 50/50 and the sample we collect has response rates of 83% from males and 75% from females, then sample weights could be calculated as:
1/.83 = 1.204819 for males
and
1/.75 = 1.333333 for females

So for a population of, say, 10,000 people, we might assume that the population actually includes approximately 5,000 males and 5,000 females.

If, for our sample, at these response rates, we get 4,150 males and 3,750 females. Adjust these using the weights to get:
4150*1.204819 = 4999.999 males
3750*1.333333 = 4999.999 females

Consequences of Ignoring Weights

A simple grade prediction model using NCES (education) data will demonstrate that differences may be observed when probabilistically sampled data are analyzed in three different ways. As a foil to this demonstration, an ordinary least squares regression model will be created and used to predict students' grades using independent variables for race, income and television.

The model used to illustrate the impact of weights is: Grades = Race, Income, TV

In the first analysis, no weights will be used. In the second analysis, the only Final Child weight will be used. In the third analysis, both the Final Child weight and the replicate weights will be used (this is the correct analysis). AM Statistical Software performs the analyses using ordinary least squares regression for the first two analyses and a Jackknife method for the last, replicated regression analysis. AM Statistical Software Beta Version 0.06.03 (c) was developed by The American Institutes for Research (A.I.R.) and Jon Cohen.

The Data

The analyses use data from the National Household Education Surveys Program of 2007, Parent and Family Involvement in Education Survey (PFI-NHES: 2007).

The model Grades = Race, Income, TV will be specified using the following variables:

1. **SEGRADES**: Overall, what are the child's grades across all subjects?

Response	Value
Mostly A's	1
Mostly B's	2
Mostly C's	3
Mostly D's	4

2. **CBLACK** – Is the child Black or African American?

Response	Value
Yes	1
No	2

3. **TVWKDYNU** – How much time does the child spend watching television or videos on a typical weekday?

Response	Value
1-16	1-16

4. **HINCOME** – What is the total household income?

Response	Value
$5,000 or less	1
$5,001-$10,000	2
$10,001-$15,000	3

$15,001-$20,000	4
$20,001-$25,000	5
$25,001-$30,000	6
$30,001-$35,000	7
$35,001-$40,000	8
$40,001-$45,000	9
$45,001-$50,000	10
$50,001-$60,000	11
$60,001-$75,000	12
$75,001-$100,000	13
Over $100,000	14

Note: the following results are generated solely for the purpose of demonstrating the differences observed when data of this type are analyzed in different ways. The reader should attempt no further interpretation of the models presented here as none of the underlying assumptions of the models has been checked.

The A.M. output for the three analyses is:

Model: SEGRADES = CBLACK HINCOME TVWKDYNU

Regression: **No Weights**

Parameter	Estimate	SE	t	p > |t|
Constant	2.860	0.081	35.120	0.000
CBLACK	-0.078	0.042	-1.879	0.060
HINCOME	-0.022	0.004	-5.436	0.000
TVWKDYNU	0.428	0.018	23.608	0.000

Regression: **Final Weight Only**

Parameter	Estimate	SE	t	p > |t|
Constant	2.913	0.118	24.611	0.000
CBLACK	-0.051	0.059	-0.863	0.388
HINCOME	-0.033	0.006	-5.942	0.000
TVWKDYNU	0.394	0.024	16.154	0.000

Replicated Regression: **All Weights**

Parameter	Estimate	SE	t	p > |t|
Constant	2.913	0.130	22.434	0.000
CBLACK	-0.051	0.063	-0.807	0.422
HINCOME	-0.033	0.005	-6.324	0.000
TVWKDYNU	0.394	0.023	16.922	0.000

AM Statistical Software Beta Version 0.06.03 (c) The American Institutes for Research and Jon Cohen

For simplicity, define the three weighting levels as:

Level 1 – No weights are used.
Level 2 – Only Final Weight is used.
Level 3 – Both Final Weight and Replicate Weights are used.

The value in the table below represents the weighting levels that produce a change when moving from the previous level. For instance, a value of "2" indicates that a change is noted when the method of analysis moves from level 1 (no weights) to level 2 (final weight only). A value of "2, 3" indicates a change for that statistics when moving among all 3 levels of weighting.

Parameter	Estimate	SE	t	p > |t|
Constant	2	2, 3	2, 3	2, 3
CBLACK	2	2, 3	2, 3	2, 3
HINCOME	2	2, 3	2, 3	2, 3
TVWKDYNU	2	2, 3	2, 3	2, 3

Summarizing, as we move from the use of no weights to the use of only the final weight to the use of both the final weight and the replicate weights (the correct analysis), the regression coefficients change with the use of the final weight, but remain the same when the replicate weights are used. This demonstrates that parameter estimates can be correctly calculated using only the final weight. The final weight alone does not, however, allow for the correct calculation of the standard error necessary for inference. This requires some form of variance estimation such as replication.

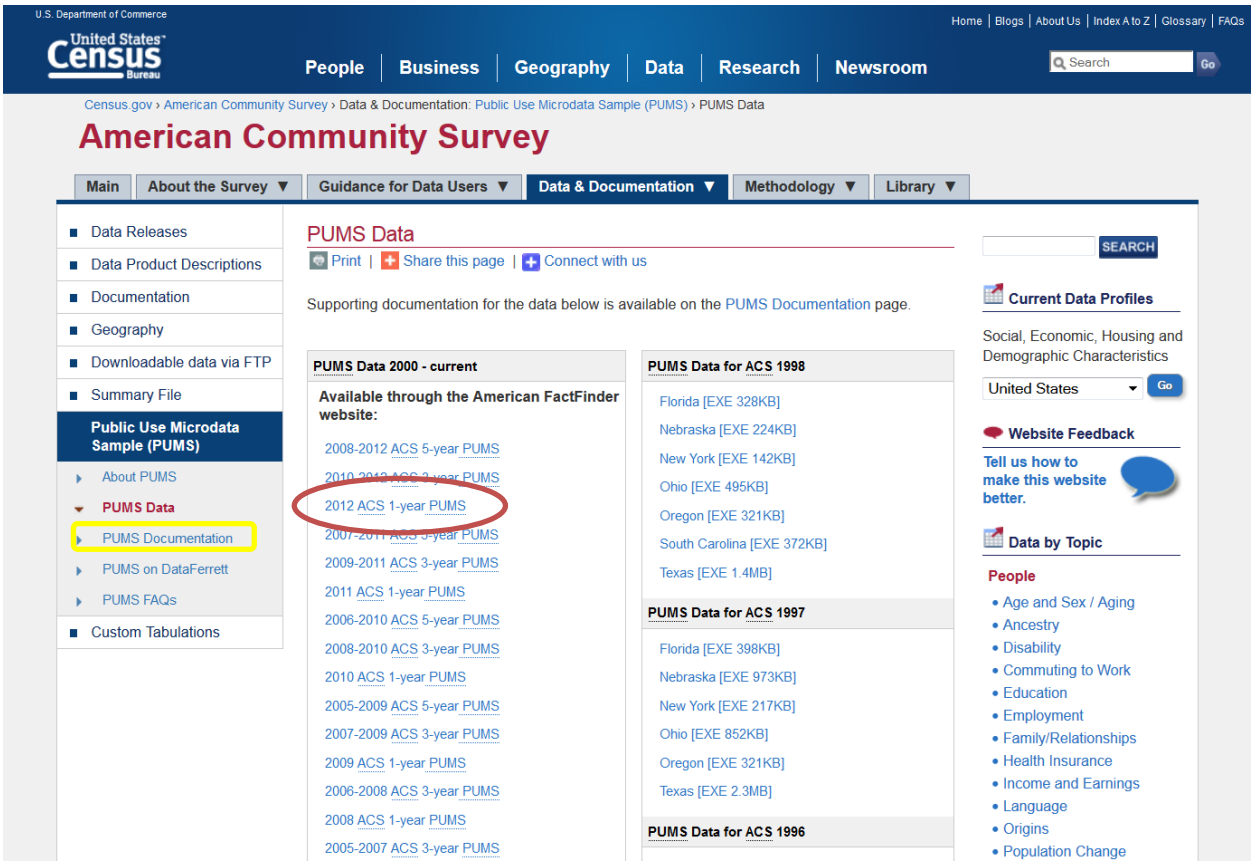
Using Data from Complex Samples

The first task with any database is to determine if weights are needed in the analysis. If weights are required, we must then determine what type of weights are involved. If the only weights required are cell weights (aka "final weight," "child weight," "household weight," "hospital weight," etc.) then the analysis can be done in SPSS or any other standard statistical package that allows weights. If, in addition to the final weight, the sample requires replicate weights (which, in my experience, are always called "replicate weights") specialized software such as SAS, Wesvar, AM, or Stata must be used.

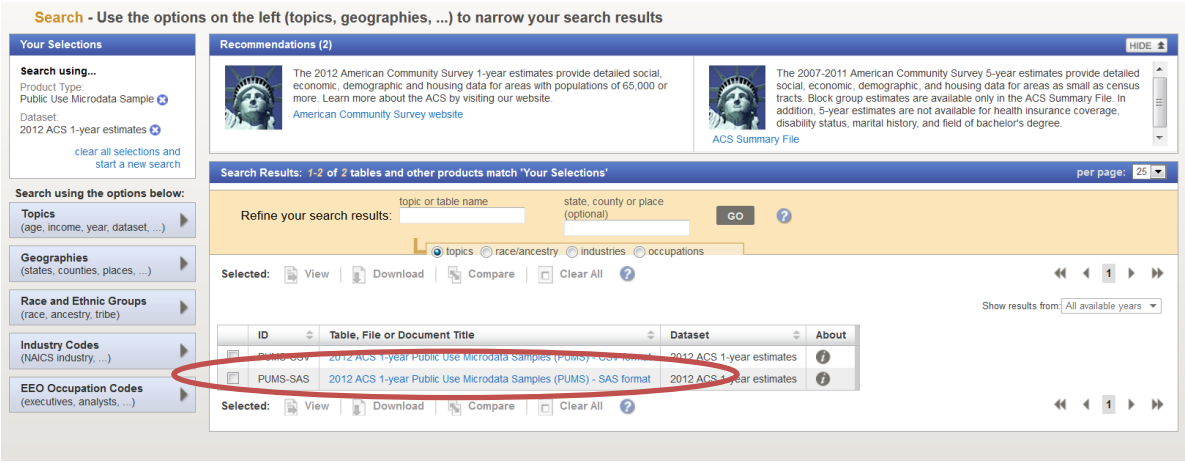
Data Analysis using PUMS Data

The P.U.M.S. data use both cell and replicate weights. To illustrate the correct use of the PUMS data and accompanying weights, we will use SAS to analyze the AGEP variable in the 2012 Alabama file by creating age categories and then constructing simple frequencies for these categories.

1) Download the data at: http://www.census.gov/acs/www/data_documentation/pums_data/. For this example, we will use the 2012 ACS 1-year P.U.M.S. data.



The data are available as either a .csv file or as a SAS dataset. We will choose the SAS dataset.

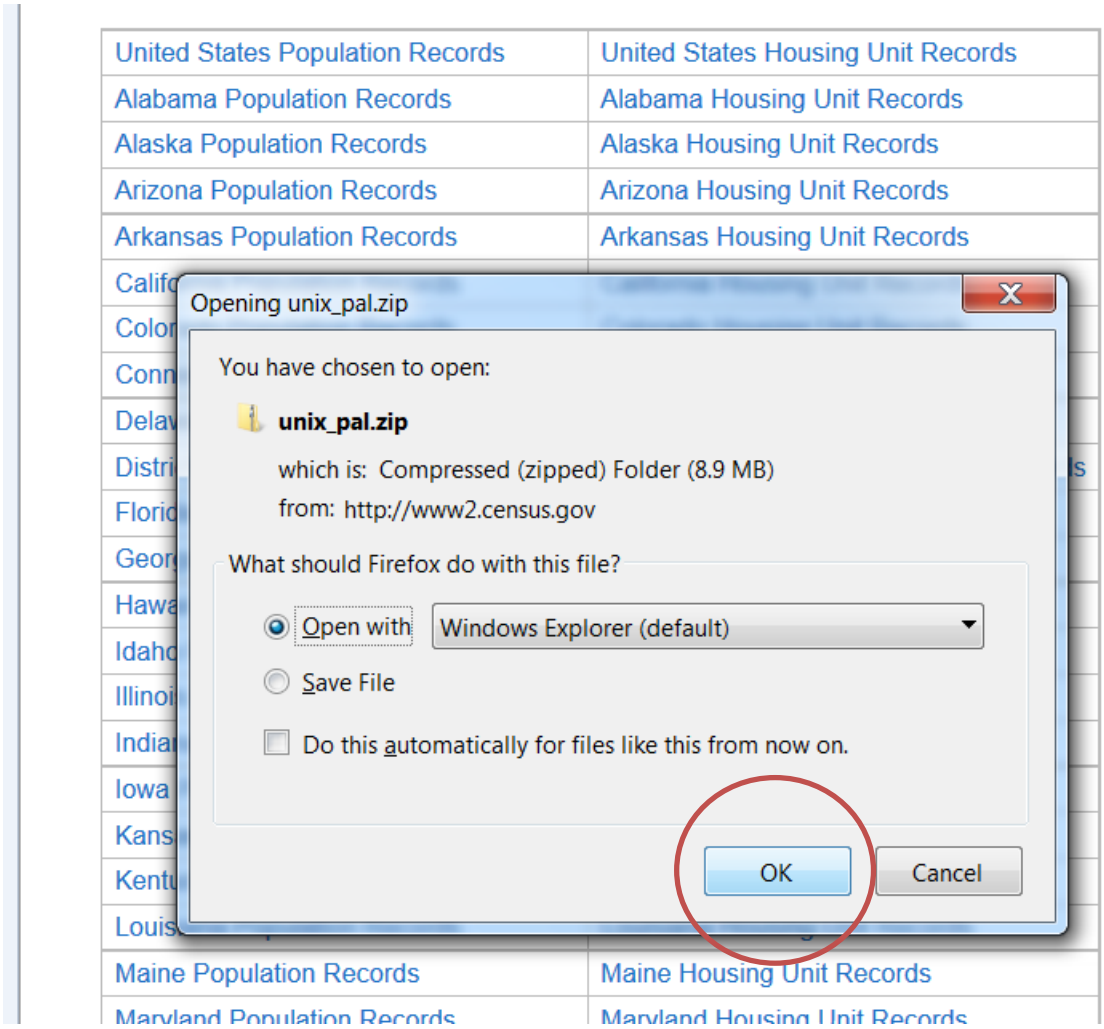


Let's work with the Alabama Population Records:

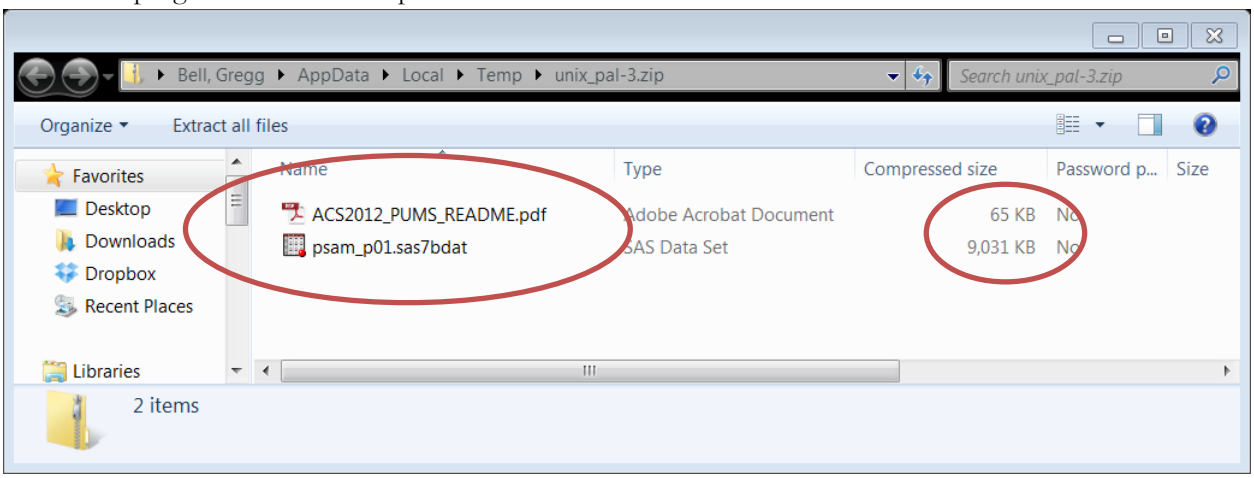
All files below are provided in SAS format. The 2012 ACS 1-year PUMS are also available in comma separated value (CSV) format.

United States Population Records	United States Housing Unit Records
Alabama Population Records	Alabama Housing Unit Records
Alaska Population Records	Alaska Housing Unit Records
Arizona Population Records	Arizona Housing Unit Records
Arkansas Population Records	Arkansas Housing Unit Records
California Population Records	California Housing Unit Records

2)



Move both of these files to an easily located folder. The act of moving the files to another folder will extract the files for most users. Notice the file sizes before and after the move. If the file is larger after the move, then it is extracted and ready to use. If the file size remains the same, then use the "Extract all files" button in this window. The README.pdf file is an important reference document. Critically, it contains instructions for using the full United States file, not covered in this tutorial. Do not open the .sas7bdat file. We will use it in our SAS program in the next step.



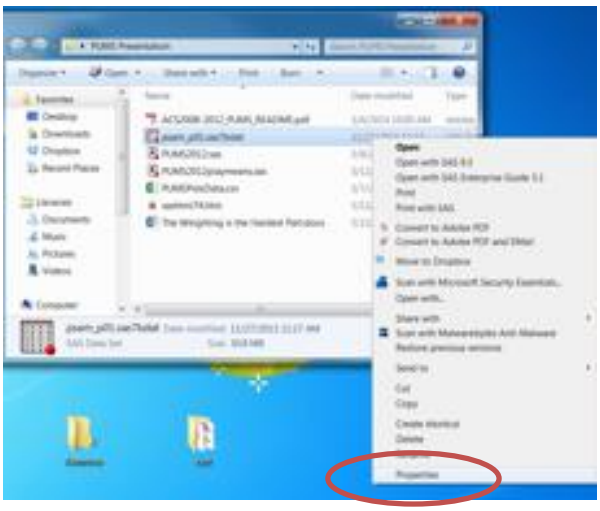
The Program

3) The SAS code below is modified version of a program on the Bowling Green State University website. The original program, along with some additional documentation can be found at: http://www2.bgsu.edu/downloads/cas/file75747.pdf.

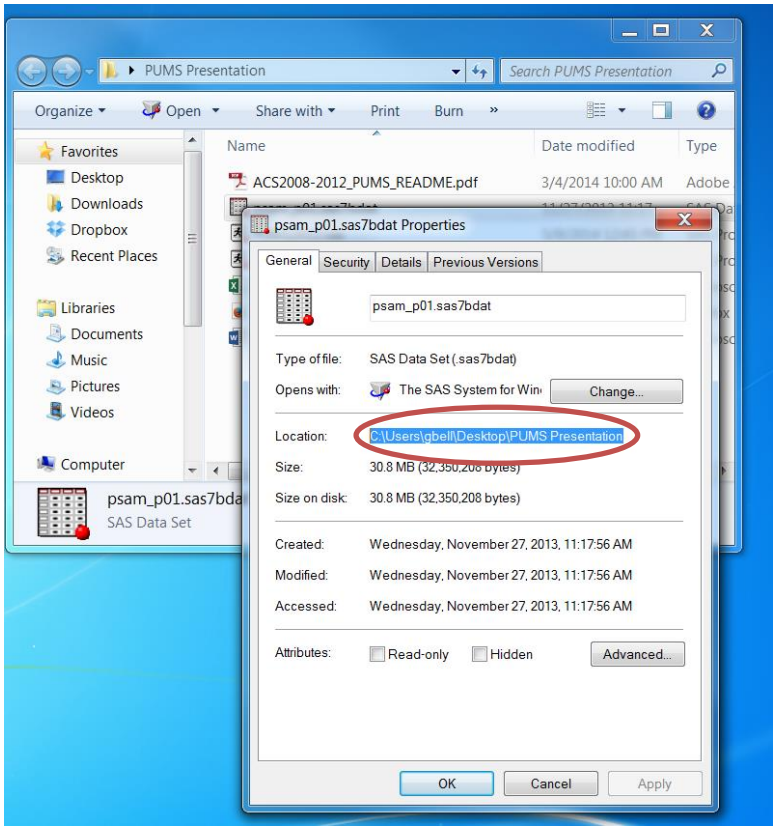
```
*1; ODS HTML CLOSE;
*2; ODS HTML;
*3; dm log 'clear' output;

*4; libname use 'C:\Users\gbell\Desktop\PUMS Presentation';

*5; data pums2012;
*6; set  use.psam_p01;
*7; if agep in(0,1,2,3,4) then agecat=1;
*8; if agep in(5,6,7,8,9) then agecat=2;
*9; run;

*10; proc sort; by agecat;
*11; run;

*12; proc surveyfreq data=pums2012;
*13; tables agecat;
*14; weight pwgtp;
*15; run;

*16; proc means data=pums2012 noprint; where agecat=1;
*17; by agecat;
*18; var pwgtp pwgtp1-pwgtp80;
*19; output out=weights sum=est rw1-rw80; run;
*20; run;

*21; data weights2 (keep=char est se cv);
*22; set  weights end=eof;
*23; if _n_=1 then sdiffsq=0;
*24; array repwts {*} est rw1-rw80;
*25; do i = 2 to 81;
*26; sdiffsq = sdiffsq + (repwts {i} - repwts{1})**2;
*27; end;
*28; if eof then do;
*29; var = (4/80)*sdiffsq;
*30; se = (var)**.5;
*31; cv = se/est;
*32; length char $20.;
*33; char = "age 0 to 4";
*34; output;
*35; end;
*36; run;

*37; proc print data=weights2;
*38; var char se;
*39; run;
```

4) To use the above program, open SAS and then copy and paste the above lines (1 through 36) into your SAS editor. Immediately save and name the program by selecting "File" and "Save As." You will need to make a few changes to the code so that it will run on your machine.

5) Begin by locating the SAS dataset (psam_p01.sas7bdat) you downloaded and extracted. Right-click on the file and select "Properties."



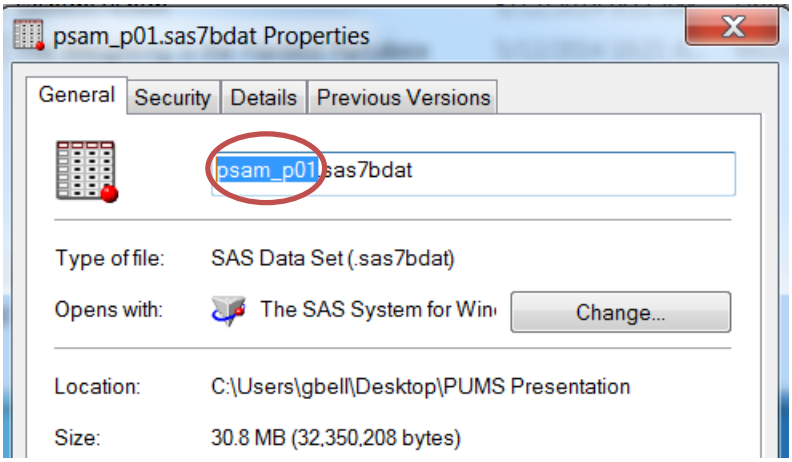
6) While depressing the left-click button on your mouse, drag over the location for the folder that contains the SAS dataset.



Paste this location between the single quotes in the libname statement (line *4; of the code.) Your location will, of course, be different from mine!

```
*3; dm log 'clear' output;

*4; libname use 'C:\Users\gbell\Desktop\PUMS Presentation';

*4; data pums2012;
*5; set  use.psam_p01;
```

7) Now copy the name of the SAS dataset from the same properties window as above.

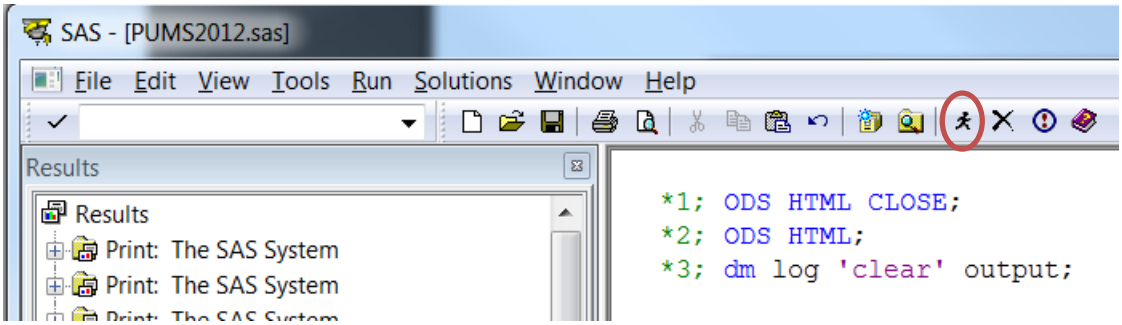


8) Paste the SAS dataset name in the data statement (line *6; of the code.) Note: you will only need to change this if you are using a different PUMS file.

```
*4; libname use 'C:\Users\gbell\Desktop\PUMS Presentation';

*5; data pums2012;
*6; set  use.psam_p01;
*7; if agep in(0,1,2,3,4) then agecat=1;
*8; if agep in(5,6,7,8,9) then agecat=2;
```

9) Under the File menu, choose "Save."

10) At this point, your program should be ready to run. To run the program, click the "running SAS programmer" button in the SAS window.



11) The results are shown below. We are primarily interested in two values on this page. The estimate (298738) and the standard error (2315.34.)

The SAS System

The SURVEYFREQ Procedure

Data Summary	
Number of Observations	47819
Sum of Weights	4822023

Table of agecat

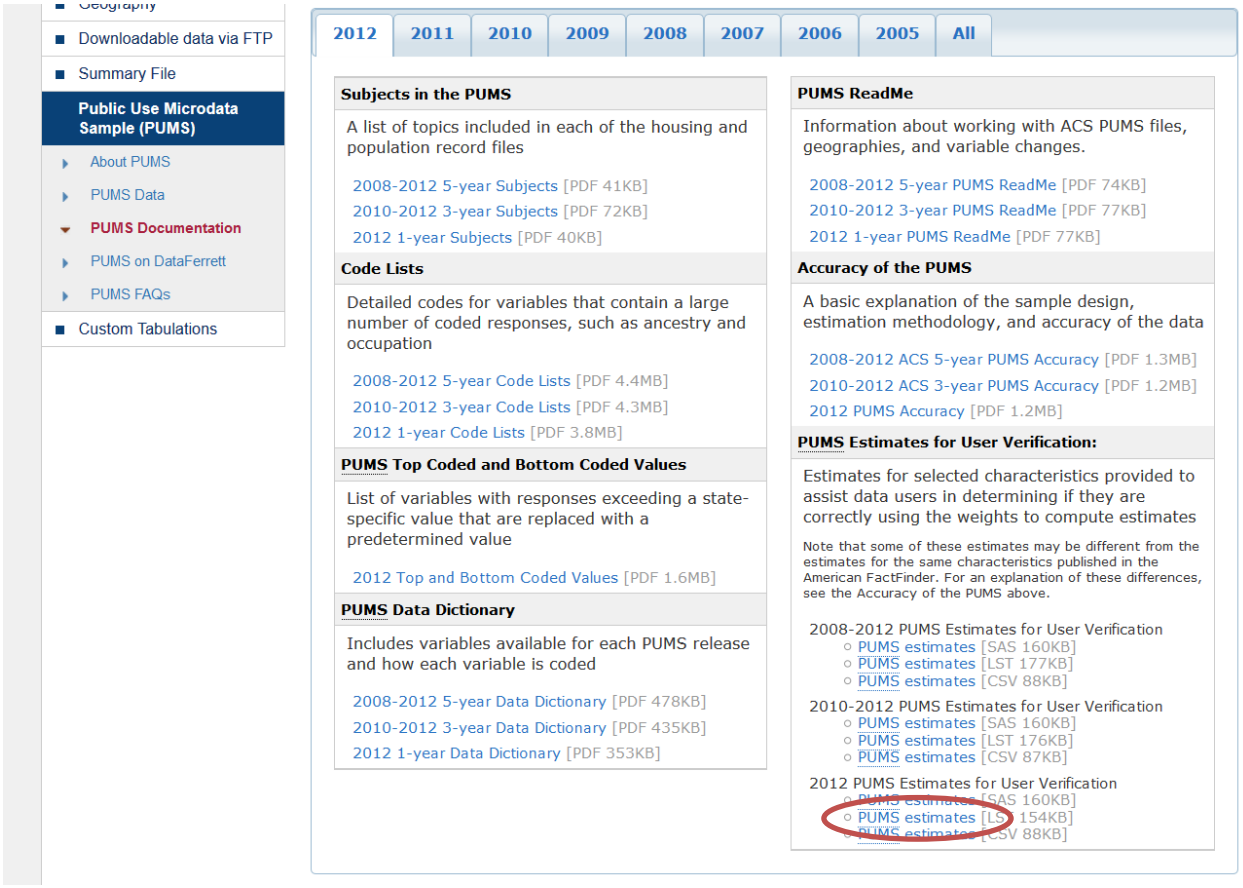
agecat	Frequency	Weighted Frequency	Std Dev of Wgt Freq	Percent	Std Err of Percent
1	2445	298738	6683	48.6416	0.9247
2	2655	315424	6893	51.3584	0.9247
Total	5100	614162	7430	100.000	

Frequency Missing = 42719

The SAS System

Obs	char	se
1	age 0 to 4	2315.34

12) It has taken a number of steps to get to this point. Fortunately, we have a way to check our numbers. Browse to:
http://www.census.gov/acs/www/data_documentation/pums_documentation/



Scroll down to Alabama to see:

State of current residence=01 State=Alabama

Characteristic	2012 PUMS Estimate	2012 PUMS SE	2012 PUMS MOE
Total population	4,822,023	0	0
Housing unit population (RELP=0-15)	4,707,169	0	0
GQ population (RELP=16-17)	114,854	0	0
GQ institutional population (RELP=16)	67,094	7	11
GQ noninstitutional population (RELP=17)	47,760	7	11
Total males (SEX=1)	2,332,059	3550	5839
Total females (SEX=2)	2,489,964	3550	5839
Age 0-4	298,738	2315	3809
Age 5-9	315,424	5890	9689
Age 10-14	322,583	5607	9224

Since these numbers match our results, our process is validated!

13) **Some brief notes about the rest of this program**:

Lines 1 through 3 simply clear the output and log so that subsequent runs of the program will not stack old upon new.

Line 4, the libname statement, assigns an internal name for the location (path and folder) of the SAS dataset we are about to use. I use the name "use," but you can change this to "bob" or "carol" or just about anything you wish. Multiple libname statements need multiple libnames, so use1, use2, etc. We only need one in this case.

Line 5, the data statement, names the SAS dataset we are about to create. The user chooses this name. I chose to call the dataset "pums2012."

Line 6, the set statement, tells SAS to create the new dataset from "psam_p01," the dataset we downloaded, which is now located in "C:\Users\gbell\Desktop\PUMS Presentation," the directory referenced in the libname statement.

Lines 7 and 8, combine data to create age categories. Category 1 is ages 0 through 4. Category 2 is ages 5 through 6. We only create the first two categories for this example.

Line 9, the run statement, tells SAS to finish the preceding step before beginning the next step. This will be true of all subsequent run statements.

Line 10 sorts the data by age categories.

Line 12 uses the Surveyfreq procedure in SAS to analyze the data. Proc Suveyfreq is a specialized procedure in SAS that correctly handles many of the needs of survey data users.

Line 13 tells proc surveyfreq to create a frequency table using the grouping variable "agecat."

Line 14 tells proc surveyfreq to use the variable "pwgtp" as the cell weights when calculating the estimate.

Line 16 begins the process of creating the standard errors. The casual user need not be concerned with the details of this process. **However**, if you changed the name the dataset in line 5, you will need to change the name after the "data=" part of this line to match the name in line 5. The where statement tells SAS to only do this for observations where agecat=1.

Line 17 tells SAS to perform the actions in line 16 for each value of the variable "agecat." That is, calculate separate estimates for agecat=1 and agecat=2. Since we are only dealing with agecat=1, this line is unnecessary. I have included here only illustrate the method for calculating estimates for additional age categories.

Line 18 tells SAS to use the weights in the dataset as variables in this part of the process. The variable "pwgptw" is the cell weight and "pwgptw1" through "pwgptw80" are the replicate weights.

Line 19 outputs these calculations to a new dataset called "weights" and names the sums of these variables "est" and "rw1, rw2… rw80."

Lines 21 through 36 use the sums created above to calculate the standard error using the formula in the technical documentation. Among other places, the formula appears in the "readme" file that accompanied the original download of the data. It is:

$$SE(X) = \sqrt{\frac{4}{80}\sum_{r=1}^{80}(X_r - X)^2}$$

Where X_r is a replicate estimate from X_1 to X_{80}, and X is the full PUMS weighted estimate.

P.U.M.S.-SPECIFIC REFERENCES

http://www.census.gov/acs/www/data_documentation/public_use_microdata_sample/

http://www.census.gov/acs/www/Downloads/handbooks/ACSPUMS.pdf

http://www.census.gov/main/www/pums.html

http://www.census.gov/acs/www/data_documentation/pums_documentation/

http://www.census.gov/prod/cen2000/doc/pums.pdf

http://www.census.gov/acs/www/Downloads/survey_methodology/acs_design_methodology_ch12_2014.pdf

http://www2.bgsu.edu/downloads/cas/file75747.pdf

GENERAL REFERENCES

AM Statistical Software. (2011). *Manual*. Retrieved July 1, 2011, from http://am.air.org/help/JSTree/MainFrame.asp

Barnett, W. S., & Belfield, C. R. (2006). Early childhood development and social mobility. *The Future of Children*, 16(2), 73-98.

Brick, J.M., Waksberg, J., Kulp, D., & Starer, A. (1995). Bias in list-assisted telephone samples. *Public Opinion Quarterly*, *59*(2), 218–235.

Brogan, D. J. (1998). *Pitfalls of using standard statistical software packages for sample survey data.* Retrieved on July 1, 2011, from http://www.hcp.med.harvard.edu/statistics/survey-soft/docs/donna_brogan.html.

Casady, R.J., & Lepkowski, J.M. (1993). Stratified telephone survey designs. *Survey Methodology*, *19*(1), 103–113.

Conway, S. (1982). *The weighting game.* Paper presented at the Market Research Society Conference, Metropole Hotel, Brighton.

Deming, W.E., & Stephan, F.F. (1940). On a least square adjustment of a sampled frequency table: When the expected marginal totals are known. *Annals of Mathematical Statistics*, *11*, 427–444.

Deming, W. E. (1943). *Statistical adjustment of data.* New York: Wiley.

Dorofeev, S., & Grant, P. (2006). *Statistics for real life sample surveys: Non-simple-random samples and weighted data*. New York: Cambridge University Press.

Efron, B. (1979). Bootstrap methods: Another look at the jackknife. *Annals of Statistics, 7*, 1-26.

Hagedorn, M., Roth, S.B., O'Donnell, K. Smith, S., & Mulligan, G. (2008). *National Household Education Surveys Program of 2007: Data File User's Manual, Volume I.* (NCES 2009-024). Washington, DC: National Center for Education Statistics, Institute of Education Sciences, U.S. Department of Education.

Hagedorn, M., O'Donnell, K., Smith, S., & Mulligan, G. (2008). *National Household Education Surveys Program of 2007: Data File User's Manual, Volume III, Parent and Family Involvement in Education Survey.*(NCES 2009-024). Washington, DC: National Center for Education Statistics, Institute of Education Sciences, U.S. Department of Education.

Hagedorn, M., Roth, S. B., Carver, P., Van de Kerckhove, W., & Smith, S. (2009). *National Household Education Surveys Program of 2007: Methodology Report.* (NCES 2009-047). Washington, DC: National Center for Education Statistics, Institute of Education Sciences, U.S. Department of Education.

Kim, J. K., Navarro, A., & Fuller, W. (2000). *Variance estimation for 2000 Census Coverage Estimates.* Proceedings of the Survey Research Methods Section, American Statistical Association, Alexandria, VA.

Kish, L. (1995). *Questions/answers from The Survey Statistician, 1978-1994.* Paris: The International Association of Survey Statisticians, Section of the International Statistical Institute.

Kish, L. (1965). *Survey sampling*. New York: John Wiley & Sons.

Kim, J, Murdock, T. & Choi, D. (2005). Investigation of parents' beliefs about readiness for kindergarten: An examination of National Household Education Survey (NHES: 93). *Educational Research Quarterly, 29*(2), 3-17.

Korn, E. L., & Graubard, B. I. (1995). Examples of differing weighted and unweighted estimates from a sample survey. *The American Statistician, 49*, 291-295.

Landis, R.J., Lepkowski, J.M., Eklund, S.A., & Stehouwer, S.A. (1982). *A statistical methodology for analyzing data from a complex survey: The first national health and nutrition examination survey (DHHS Pub. No. 82-1366).* Vital and Health Statistics, Series 2, No. 92. Washington, DC: National Center for Health Statistics.

Moser, C. A., & Kalton, G. (1971). Survey methods in social investigation (2nd ed.). London: Heinemann Educational Books.

Quenouille, M. (1949). Approximation tests of correlation in time series. *Journal of the Royal Statistical Society B, 11*, 18-84.

Shao, J., & Tu, D. (1995). *The jackknife and bootstrap, Springer series in statistics.* New York: Springer-Verlag.

Sharot, T. (1986,) Weighting survey results. *Journal of Market Research Society, 28*(3), 269-284.

Wild, C. J., Seber, & George, A. F. (2000). *Chance encounters: A first course in data analysis and inference.* New York: John Wiley & Sons.