cepr
CENTER FOR ECONOMIC AND POLICY RESEARCH

SIPP User Notes

To: SIPP Users
From: Heather Boushey, economist, Center for Economic and Policy Research
RE: Set A: ID's and Weights
Date: December 22, 2005

Set A includes identification variables and sample weights for national estimates. The SIPP is a nationally representative survey at the household level and is not designed specifically to generate estimates at the sub-national level. However, the U.S. Census has recently provided experimental weights for use in conducting state-level analysis. The use of these weights is addressed in this Memo.

ID

CEPR's extraction from the raw Census data creates a variable *id* that uniquely identifies each individual in the SIPP. The *id* variable is created from the identifying variables (with the 1993 panel and earlier names in parentheses), *ssuid* (*suid*), *eentaid* (*entry*), and *epppnum* (*pnum*). (See *SIPP Users Guide*, chapter 10.)

For 1993 and earlier, analysis should only be done on observations where the variable indicating "in sample," *pp_mis*, is equal to 1. This is the only reliable guide for whether or not to include an individual. (See *SIPP User's Guide*, p. 9-5.)

Variance stratum code

The variance stratum codes are for use with statistical packages that allow the programmer to adjust the standard errors. However, this may need to be done manually. (See the *SIPP 1996 Panel*, chapter 8, for more details.)

Person, Family, and Household Weights

Each household and each person within each household has four weights: *wffinwgt* (family weight for the reference month), *whfnwgt* (household weight for the reference month), *wpfinwgt* (person weight for the reference month), and *wsfinwgt* (related sub-family weight for the reference month). For most purposes, the person weight (*wpfinwgt*) should be used. When using the family or household weights, the user will have to verify that definitions of "family" or "household" are the same as the definitions used to generate the family and household weights. If you are using CEPR's Set C data (Household and Family relationships), these correspond to families defined by the variable *hhid* for households, *pfid* for families, and *sfid* for sub-families. (See *SIPP User's Guide*, pp. 8-10 - 8-13.) Longitudinal person weights for specific calendar years are also included for the 1992 and 1993 panels: *fnlwgt92*, *fnlwgt93*, and *fnlwgt94*. (See *SIPP User's Guide*, p. 8-16.)

The four primary weights are generated for each reference month during the panel and they take into account survey attrition. The weights can be averaged to form estimates of monthly averages over some period of time. For example, one can estimate the monthly average number of households in a specified income range over November and December 1996 using the household weight for those months. However, it should be noted that there is no weight for characteristics that involve a person's or household's status over two or more months (such as, number of households with a 50 percent increase in income between November and December 1995).

Sub-national analysis

Analysis of metropolitan areas or regions requires the use of an adjustment factor. (See the *SIPP96 longitudinal codebook*, p. 8-4 for more information about this issue.)

State Weights

Set A also includes the U.S. Census Bureau's experimental state weights (*statewgt*) for the 1996 panel. These state weights enable SIPP users to generate state-level estimates with the same degree of sampling accuracy as in the national sample. However, the U.S. Census Bureau recommends that only the largest 5-10 states have a sufficient number of observations to produce stable statistical results consistent with random samples within a particular state. As a rule of thumb, users might consider using state weights only for panel samples with over 100,000 person-months for a particular state. This includes the following 11 states: CA, FL, GA, IL, MI, NJ, NY, NC, OH, PA, and TX.

The file *StateWeightCheck.xls* (available on this cd) compares SIPP means and medians for race/ethnicity and income, using both the person weights and the state weights, to estimates generated from the CPS, which is designed to be representative at the state level. When the SIPP sample size for a particular state is small, the accuracy of even a simple mean often varies widely from the CPS estimate. For example, in the District of Columbia (DC) or Alaska, where person-months in the SIPP are generally less than 10,000 per year, the SIPP and CPS estimates of the percent of the population that is white (or the median income estimate) diverge more so than in states such as New York or Texas, where the SIPP has more than 100,000 person-months for any particular year.

While it should be intuitive to understand why larger SIPP sample sizes yield summary statistics that are closer to the CPS estimates of the same parameter, we can check this empirically by regressing the absolute value of the difference between the SIPP estimate and the CPS estimate on the (log of the) number of observations in the SIPP sample, while controlling for each year of the panel. The absolute value of the differences between the SIPP and CPS estimates are calculated for the mean value of the percent white and the median income level, with the SIPP estimates using both the person weights and the state weights.

We estimate the following model:

$$Dependent\ Variable = \beta_0 + ln(obs) + yr96 + yr97 + yr98$$

where the respective dependent variables are the absolute value of the (percentage) difference between the CPS estimate and the SIPP estimate:

abs [(SIPP median income (person weight) - CPS median income)/SIPP median income (person weight))]

abs [(SIPP median income (state weight) - CPS median income)/SIPP median income (state weight))]
abs [(SIPP percent white (person weight) - CPS percent white)/SIPP percent white (person weight))]
abs [(SIPP percent white (state weight) - CPS percent white)/SIPP percent white (state weight))]

yr96, yr97, yr98 and *yr99* (dropped) are year dummies; and *ln(obs)* is the log of the number of observations for a particular state in a given year.

We expect that if larger state sample sizes (states with more observations) are associated with smaller differences between the CPS estimates and the SIPP estimates, then the coefficient for the log of the number of observations should be negative and highly significant.

The difference correlations in Table 1 show that larger state samples are associated with smaller differences between the state-level SIPP and CPS estimates. The same is true for regions, but only in the "percent white" models. The region models should be viewed with caution, however, because the number of observations within each region is relatively large to begin with and the sample size for the regression in question (N=36) is probably too small to yield definitive results.

Table 1. Difference Correlations

Observations	Dependent Variable	Ln(obs): β	t-statistic
State model			
184	median income (person weight)	-.0158	-2.85
184	median income (state weight)	-.0182	-3.75
184	percent white (person weight)	-.0409	-4.92
184	percent white (state weight)	-.0405	-6.14
Region model			
36	median income (person weight)	-.0041	-0.27
36	median income (state weight)	.0100	0.52
36	percent white (person weight)	-.0130	-4.27
36	percent white (state weight)	-.0094	-2.76

Note: We use White robust standard errors. Regressions with N=184 include states from the years 1996, 1997, 1998, and 1999 using the SIPP 1996 panel. Regressions with N=36 are for the regions for the same years. States that are combined in the SIPP data—for example Maine and Vermont—are not included in the analysis.

Table 2 shows the mean values for the variables in question—the two SIPP estimates and the CPS estimate. For "percent white" and median income, the SIPP estimates using the state weights are closer to the CPS estimates than the SIPP estimates using the person weights, though these differences are still statistically significant.

One question we have not yet answered, but is certainly worth exploring, is whether or not (and how) researchers can group states together to gather enough observations to yield a random sample and hence stable statistical results.

Table 2. Difference of Means: States

Variable	Observations	Mean	Standard error	90% Confidence interval		t-stat (diff<0)
Percent White						
SIPP (person wgt)	184	.745	.0128	.724	.766	
SIPP (state wgt)	184	.755	.0122	.735	.776	
CPS	184	.762	.0114	.743	.781	
Difference						
SIPP (person wgt) – CPS	184	-.0170	.0044	-.024	-.010	-3.884
SIPP (state wgt) – CPS	184	-.0064	.0036	-.012	-.000	-1.679
Median Income						
SIPP (person wgt)	184	2924	44	2850	2997	
SIPP (state wgt)	184	2956	45	2882	3030	
CPS	184	3171	37	3110	3232	
Difference						
SIPP (person wgt) – CPS	184	-247	25	-288	-206	-10.02
SIPP (state wgt) – CPS	184	-214	23	-253	-176	-9.16

Note: The analysis for Table 2 is for the years 1996, 1997, 1998 and 1999, using the 1996 SIPP panel. States that are combined in the SIPP data—for example Maine and Vermont—are not included in the analysis. These means are not weighted by size of state and cannot be compared to other aggregate estimates for the time period in question.

References

U.S. Department of Commerce, Economic and Statistics Administration. U.S. Census Bureau. 2001. *Survey of Income and Program Participation Users' Guide*. 3rd Edition. Washington, DC.

U.S. Department of Commerce, Economic and Statistics Administration. U.S. Census Bureau. *Survey of Income and Program Participation (SIPP) 1996 Panel, Longitudinally Edited Waves 1-12 Person-Month Microdata Files Technical Documentation*.

Appendix 1: U.S. Census Bureau documentation on experimental state weights

Notes to Users of the SIPP 96 Panel state-based weights:

These weights are "research weights." They have not been independently verified. While the SIPP branch has done some analysis of the weights, and has determined that the weights do not seriously deviate from what we expected, no one outside the Branch has examined them. As such, we would like users to provide us with feedback.

We have also created GVF parameters for each state, excluding the non-disclosure states, Vermont, North Dakota, Maine, South Dakota, and Wyoming. These GVFs are based on wave 2 data and should be applicable for waves 1 through 5. At a later date we will distribute additional parameters.

Soon we will complete a more complete technical paper providing more details on using these weights.

Some caveats:

1) These are research weights. They have not gone through the normal process of verification and validation. They may have errors in them.

2) Some states have very few sample cases and thus will not produce accurate or stable results, e.g., DC, Montana, Alaska.

3) Only the 5 to 10 largest states will provide estimates with stability and accuracy similar to national estimates.

4) Totals by state or at the national levels are not comparable to the same totals produced by the national based weights because the population controls are different. The controls are based on state level estimates.

This CD contains 12 files of weights, one for each wave of the 1996 panel of SIPP. These person month files are with state based final weights. The SAS input format is:

@1 STPMID $ 12. @14 STPMMON 1. @16 STPMRFRP 1. @19 STPMPNUM 4. @25 STPMFWGT 12.4

The variables are: scrambled public use id, reference month, family head indicator, person number, and final weight. The family head indicator is "1" if a person is a family head (either family or sub-family) and a ".", "SAS missing value code," if a person is not a family head. The file may be read as space delimited or with a format statement. In order to ensure proper merging with other SIPP data files the STPMID must read ? as a character variable.

If you have any questions contact John Boies at (301) 763-5923, Steve Mack at (301) 763-4182, or Tracy Mattingly at (301) 763-1919.

Appendix 2: Variable List

Variable	Label
Merging Variables (appear on every Set)	
id	Unique ID
srefmon	Reference month
wave	Wave of data collection
age	Age in this month
Variables in this set only	
eentaid	Person's interview status for this interview
eppintvw	Edited person number
epppnum	Person number
fnlwgt90	Person's weight assigned for 1990
fnlwgt91	Person's weight assigned for 1991
fnlwgt92	Person's weight assigned for 1992
fnlwgt93	Person's weight assigned for 1993
fnlwgt94	Person's weight assigned for 1994
gvarstr	Stratum code for variance estimation
lgtcy1wt	Longitudinal first calendar year
lgtcy2wt	Longitudinal second calendar year
lgtcy3wt	Longitudinal third calendar year
lgtkey	Person longitudinal key
lgtpnwt1	Longitudinal panel weights
lgtpnwt2	Longitudinal panel weights
lgtpnwt3	Longitudinal panel weights
month	Calendar month of the reference
panel	Panel year
pnlwgt	Panel person's weight assigned
pp_mis	Person's interview status for this month
rot	Rotation group number
shhadid	Address ID
ssuid	Sample unit identifier (string)
ssuseq	Sequence number of person
statewgt	State weight
wffinwgt	Family weight for the reference month
whfnwgt	Household weight for the reference month
wpfinwgt	Person weight for the reference month
wsfinwgt	Related subfamily weight for the reference month
year	Calendar year