# SIPP: LONGITUDINAL ESTIMATION FOR PERSONS' CHARACTERISTICS

Edward L. Kobilarcik and Rajendra P. Singh, U.S. Bureau of the Census

## I. INTRODUCTION

The Bureau of the Census has been conducting interviews for the Survey of Income and Program Participation (SIPP) since October 1983. The SIPP is a national survey and is designed to provide improved information on the income and participation in government programs of the noninstitutional United States population. Person and household characteristics that may influence income and program participation are also available from the SIPP. This information is vital to improve the capability of federal agencies to formulate and evaluate their policies and programs in the areas of income and social welfare.

Two types of estimates will be produced from the survey--cross-sectional and longitudinal. The method developed for producing cross-sectional estimates is described in [5]. This paper presents estimation methodology to provide longitudinal estimates of person characteristics from SIPP data. We define longitudinal estimates to be those that are obtained by linking two or more interview data files. These estimates include the length of time in a particular state (spell estimate), transition estimates at any given time or interval, annual estimates of income and estimates of change of certain characteristics. The method presented in this paper is developed for the first SIPP longitudinal file covering the first three interviews of the survey. This method consists of several stages of weight adjustments designed to reduce the bias in the survey caused by undercoverage and nonresponse. (These estimation stages do not differ appreciably from those used in SIPP cross-sectional estimation.)

This file has been developed to be used primarily for research purposes and the estimation method may be revised for future longitudinal products. Because of the urgency to make this file available for the summer of 1986, some of the decisions concerning the estimation method may not be conceptually sound, for example, treating those households in which at least one household member failed to respond to the first interview as a nonresponding household. However, the increase in bias and/or variance due to these decisions is expected to be negligible.

## II. BACKGROUND AND SAMPLE DESIGN

The SIPP 84 panel is a multistage stratified systematic sample of the noninstitutionalized resident population of the United States. This population includes persons living in group quarters, such as dormitories, rooming houses, and religious group dwellings. Noncitizens of the United States who work or attend school in this country and their families were eligible. All other persons were ineligible. This includes crew members of merchant vessels, Armed Forces personnel living in military barracks, and institutionalized persons, such as correctional facility inmates and nursing home residents. With these qualifications, persons who were residing in the United States at the time of the first interview were eligible for SIPP. However, only persons who were at least 15 years of age were eligible for interview.

Initially, a sample of living quarters in 174 Primary Sampling Units (PSUs) was selected. (Living quarters are those in which the occupants do not live and eat with any other person in the structure and that have either direct access from the outside of the building or through a common hall, or complete kitchen facilities for that unit only.) These 174 PSUs were subsampled from the Current Surveys (CS) A design PSUs [1]. To subsample these PSUs, SIPP strata were formed by combining CS strata having sample PSUs with similar proportion of non-white persons (1970), urban persons (1970), and families with income below the poverty level (1969). Forty-five of the CS strata were single-PSU strata and were selected in SIPP with certainty. To select the remaining 129 nonself-representing (NSR) PSUs, a CS stratum was selected from each SIPP stratum with probability proportional to its size. The CS PSUs in the selected CS strata were the designated NSR sample SIPP PSUs.

The SIPP sample is divided into four groups of equal size called rotation groups. One rotation group is interviewed each month. In general, one cycle of four interviews is called a wave. This design provides a smooth and steady work load for data collection and processing. Persons in the sample are interviewed once every four months for approximately two and one-half years. The reference period for the interview questions is the four months preceding the interview month. For example, the reference period for the November 1983 interview month is July through October 1983. These sample persons are interviewed again in March 1984 for the November 1983 through February 1984 period.

Persons 15 years old and over present as household members at the time of

first interview are to be part of the survey for the entire two and one-half year period. With certain restrictions, these sample persons are followed if they move to a new address. "New" persons living with sample persons are considered to be part of the sample only while residing with these sample persons. More details on the SIPP design are given in [2], [3], and [4].

## III. LONGITUDINAL UNIVERSE OF PERSONS

Before defining the longitudinal universe, it is first necessary to consider the SIPP universe at the beginning of the survey and the possible ways persons can enter and exit this universe.

As mentioned previously, the SIPP universe at the beginning of the survey is persons who are members of the civilian non-institutional population and members of the military not residing in military barracks. Persons can enter the SIPP universe in two ways: 1) persons can move from foreign living quarters, institutions or military barracks (call these places ineligible addresses) to an eligible address or 2) persons can be born to members of the universe. Likewise, persons can exit the universe in two ways: 1) moving to an ineligible address or 2) dying. A more comprehensive discussion is presented in [9].

With the above in mind, the longitudinal universe is defined to be the noninstitutional population (excluding military barracks) on December 1, 1983. This date is the midpoint of the wave 1 interview months. With this definition, the sample from the universe is restricted to only those persons who were eligible for the first SIPP interview. Because of this, persons who relocate from an ineligible address to an eligible one during the time period after the first interview are excluded from the universe since they were not in the eligible population during the first interview. However, eligible persons who die or move to an ineligible address are included since they were in the eligible population.

## IV. SAMPLE OF UNIVERSE

The sample from the longitudinal universe consists of eligible persons living in the selected living quarters at the time of the first interview. Not all of these persons were interviewed. Those who did respond to the initial interview are called original sample persons. Longitudinal analysis will only be appropriate for these original sample persons. This sample can be viewed as a sample of cohorts, with the cohorts being those persons in the SIPP sample between October 1983 and January 1984, inclusive. (By definition, a cohort is a group of individuals sharing a common characteristic.) Longitudinal analysis will only be appropriate for these cohorts. Longitudinal tabulations for such persons can be produced for the time period covered by the first 12 interview months of the survey. This corresponds to the period October 1983-September 1984 and represents the first three SIPP interviews. Data for 12 months will be available for each original sample person except for those who are known to have left the universe. The specific 12 months available depends on the person's rotation. For example, rotation 1 has data for reference months June 1983-May 1984.

## V. OVERVIEW OF LONGITUDINAL ESTIMATION

Three strategies are generally suggested as solutions to handle whole interview nonresponse - a weighting adjustment, imputation, or a combination of the two [10], [11]. In our estimation procedure, all such cases will be handled by a weighting adjustment. We decided on this approach because of time and resource constraints and the unavailability of a good longitudinal imputation system.

The estimation procedure defined below is used to develop longitudinal weights for original sample persons. Certain processes in the procedure were developed to reduce for some, but not all, of the known biases in the SIPP such as bias due to undercoverage and attrition. These biases are briefly discussed in VII.

A ratio estimation technique is used in the longitudinal estimation. A set of variables correlated to estimates of interest is used to define ratio adjustment cells for various adjustments. For a given cell, the ratio adjustment factor for each respondent in that cell $R_c$ is obtained as

$$R_c = \frac{T_c}{W_e}$$

, where $T_c$ is a control total and $W_c$ are the weighted counts that are adjusted to the control total. The control total may be obtained from the sample or from an independent source.

The following three assumptions are implicit in the formation of these cells:

1. There is a significant correlation between the important survey estimates and the variables used to form weighting cells.
2. Two different weighting cells have different means.
3. Within each weighting cell, the means for the sample respondents and the nonrespondents are equal.

Thus, it is desirable to form a new cell if the mean of characteristics of interest and the response rate for this

cell are different from the mean and response rate from all other weighting cells.

Each sample person is given a single longitudinal weight, with this weight being assigned to each of a person's 12 reference months.

## VI. DESCRIPTION OF ESTIMATION PROCEDURE

Several processes are included in the construction of longitudinal weights. Each process has its own specific objective. As explained below, the processes consist of several adjustments.

The following sample persons will be treated as "interviewed" persons in the estimation procedure: 1) those who responded to each of the first three interviews and who during the first interview lived in a household in which all eligible members responded to the interview (call this a wave 1 interviewed household) and 2) those who resided in a wave 1 interviewed household but who during the time period covered by the second and third interview are known to have died or moved to an ineligible address (foreign living quarters, institutions or military barracks). For persons who are known to have died or moved to an ineligible address, the months that such persons were deceased or residing in an ineligible address will be identified.

The following sample persons will be treated as "noninterviewed" persons in the estimation procedure: 1) those who at the time of the first interview lived in a household in which at least one household member failed to respond to the first interview (call this a wave 1 noninterviewed household), 2) those who resided in a wave 1 interviewed household but failed to respond to the second and/or third interview because of household or person nonresponse, and 3) those who resided in a wave 1 interviewed household but who moved in with members of another wave 1 interviewed household after the first interview. (This occurred for only four households.) These persons are treated as noninterviews because an imputation system for handling missing interviews is not yet available and because the processing system is unable to handle households defined in 1) and 3) above.

All persons classified as interviewed are assigned positive weights, while those classified as noninterviewed are assigned zero weights.

### B. Preparation of Unbiased Estimates

A common method of estimation, weighting by the reciprocal of the probability of selection ($P_i$), is the first step in the weighting process. This procedure results in an unbiased estimator of a population total assuming 100% response. With this procedure, the unbiased weight for the $i^{th}$ sample person is

$$W_i = \frac{1}{P_i}$$

where $P_i$ is the selection probability of the household containing the $i^{th}$ sample person. For some households a factor is included in the selection probability because different overall sampling fractions were used for certain parts of the population. In particular, certain units were subsampled because their actual size was much larger than anticipated.

### C. Adjustment for Noninterviews

The next step in the estimation process is the adjustment for noninterviews. In general, noninterview weight adjustment consists of the reassignment of the weights of noninterviewed households or persons to groups of interviewed households or persons that hopefully have similar characteristics. This is equivalent to assigning the mean value of the cell to all nonrespondents. Noninterview adjustment will take place in two phases. The first phase consists of a household adjustment, while the second phase is a person adjustment.

In the first phase, a household adjustment is made to account for persons who resided in a wave 1 noninterviewed household. The adjustment consists of the computation of weight adjustment factors within cells defined by cross-classifications of the following variables:

1. Census region (Northeast, Midwest, South, West)
2. Residence (Metropolitan, non-Metropolitan)
3. Race of reference person (Black, non-Black)
4. Tenure (owner, renter)
5. Household size (1, 2, 3, 4 or more)

The cell assigned to each household is based on the values of these variables as of the initial SIPP interview.

The second phase of the adjustment accounts for persons who resided in a wave 1 interviewed household but who failed to respond to at least one of the remaining two interviews for reasons other than death or moving to an ineligible address. The adjustment is on a person basis and consists of the computation of weight adjustment factors within cells defined by cross-classifications of the following variables:

1. Average monthly household income of the person's household
2. Program participation status of person's household
3. Person's labor force status
4. Person's race and ethnicity

5. Years of school completed by person
6. Assets ownership status of person's household

For each of the two adjustment phases, the following ratio is computed within each noninterview adjustment cell using the weighted counts of households (or persons):

$$1 + \frac{\text{Noninterviewed households (persons)}}{\text{Interviewed households (persons)}}$$

(For variance considerations, individual cells are combined together if the ratio in a given cell is too large or contains too few cases.) For a given cell, this ratio is $F_{1c}$ for the household adjustment phase and $F_{2c'}$ for the person phase. These ratios are applied to the initial weight $W_i$ of each interviewed person within a given noninterview cell. At the completion of the noninterview adjustment procedure, each person bears a weight equal to the following product:

$$W_i \times F_{1c} \times F_{2c'}$$

After the noninterview adjustments are made, noninterviewed persons are assigned zero weights. Further processing is limited to interviewed persons.

C. Adjustments To Demographic Differences From Total Population

The weighted distribution of the sample generally differs somewhat from the distribution of the total population with respect to demographic variables. This is due to two reasons. First, the distribution of the sample PSUs may not accurately represent the distribution of all PSUs due to sampling errors. This arises because in some areas one PSU is selected to represent an entire stratum of PSUs. Secondly, there exists undercoverage of households and persons within these households.

In order to reduce the mean square error (MSE) of survey estimates, two stages of adjustment are used to help bring the weighted sample distribution and the population distribution into closer agreement. This is accomplished by post-stratifying using demographic variables that are highly correlated with the variables to be measured. The first stage is designed to adjust for the sampling error associated with the sample PSUs. Undercoverage is adjusted in the second stage. Both stages are explained in greater detail below.

1. First Stage Adjustment

First stage adjustment employs a cell by cell weight adjustment procedure applied to households. For various categories of race and residence defined by the variables specified below, ratios were calculated within each adjustment cell reflecting the relationship between the estimated 1980 census household counts generated from the SIPP sample to the total population at the time of the 1980 census. (Adjustment cells are collapsed if the ratio in a given cell is too large or contains too few cases.)

a. Census region
b. Residence
c. Central city status
d. Race of household head

The weight after this adjustment is called the "first-stage weight" and is equal to the following product:

$$W_i \times F_{1c} \times F_{2c'} \times \text{(First-stage ratio)}$$

2. Second Stage Adjustment

The second stage of adjustment is applied to interviewed persons to account for undercoverage by bringing the distribution of sample persons into closer agreement with independently derived current estimates. These independent estimates are obtained using a Current Population Survey (CPS) estimation procedure developed for the CPS March income supplement [5]. The CPS estimates are used because they have a lower variance than SIPP estimates. This in turn increases the precision of the SIPP estimates.

Separate procedures are applied to sample persons aged 14 and under (children) and sample persons age 15 and over (adults). For children, a cell by cell adjustment is applied in several race x age x sex cells. For adults, a "raking" procedure is applied to adjustment tables defined by the following variables: race, age, sex, householder status, and relationship to householder status. A cell by cell adjustment for Hispanics is applied to both children and adults.

a. General Description

1. Raking Procedure for Adults

In brief, the "raking" procedure is an iterative weight adjustment procedure which aligns weighted sample counts with known marginal distributions. The method of iterative proportions which provides a best asymptotic normal (BAN) estimator in [7] is used. The procedure is used in our weighting process as one part of the second-stage adjustment for persons aged 15 years and over. It is applied here to the first-stage ratio estimates of these persons.

2. Description of Raking Procedure

The raking procedure defined below is for two marginal distributions. Define:

$W_{ijk}$ = first-stage weight of $k^{th}$ person in $i^{th}$ row, $j^{th}$ column.

$Y_{i.}$ = CPS control estimate for $i^{th}$ row.

$Y_{.j}$ = CPS control estimate for $j^{th}$ column.

We wish to obtain adjusted first-stage weights $W_{ijk}$ such that $\sum_{jk} W_{ijk} = Y_{i.}$ and $\sum_{ik} W_{ijk} = Y_{.j}$

The above is accomplished by applying the Deming-Stephan method of deriving the $W_{ijk}$ by proportionally adjusting the interior cell values until in turn each of the marginal equations is satisfied [7]. Each adjustment begins with the outcome of the previous adjustment. The process is completed when the condition equations are satisfied to a specified tolerance.

The procedure is conducted in the following manner. Below, all row adjustments are labelled with an odd superscript, while column adjustments are given an even superscript.

First, a ratio adjustment factor for the rows is computed as

$$f_{i.}^{(1)} = Y_{i.} / \sum_{j} \sum_{k} W_{ijk} \quad ,$$

followed by the computation of a column ratio adjustment factor:

$$f_{.j}^{(2)} = Y_{.j} / \sum_{i} \sum_{k} W_{ijk} \, f_{i.}^{(1)} \quad ,$$

followed by the computations of another row ratio adjustment factor

$$f_{i.}^{(3)} = Y_{i.} / \sum_{j} \sum_{k} W_{ijk} \, f_{i.}^{(1)} \, f_{.j}^{(2)} .$$

Then, an estimate of the column marginals after the third iteration is computed:

$$\hat{Y}_{.j}^{(3)} = \sum_{i} \sum_{k} W_{ijk} \, f_{i.}^{(1)} \, f_{.j}^{(2)} \, f_{i.}^{(3)}$$

If $| \hat{Y}_{.j} - Y_{.j}^{(3)} |$ T for all j

(where T is some defined level of tolerance), then the procedure is terminated and each interior cell is assigned an overall ratio adjustment factor computed as

$$g_{ij} = f_{i.}^{(1)} \, f_{.j}^{(2)} \, f_{i.}^{(3)} .$$

If the tolerance is not met, the process is continued. After each odd iteration

$| Y_{.j} - \hat{Y}_{.j}^{(Z)} |$ is checked to see if the tolerance is met for all j. The procedure is terminated when all columns meet the specified tolerance. If the process is terminated after z iterations, each cell is then assigned a ratio adjustment factor

$$g_{ij} = f_{i.}^{(1)} \, f_{.j}^{(2)} \ldots f_{.j}^{(Z-1)} \, f_{i.}^{(z)} .$$

3. Adjustment for Hispanics

Part of the overall second-stage procedure consists of a Hispanic adjustment procedure in order to reduce the MSE of SIPP Hispanic estimates.

For various sex by age categories of the Hispanic population, ratios are calculated based on the relationship between weighted Hispanic estimates and independent Hispanic estimates. The ratio is applied only to Hispanic persons.

4. Age Adjustment for Children

For persons 14 years of age and under, an age adjustment procedure is applied to reduce the MSE of children estimates.

For such persons, ratios are computed based on the weighted estimates of children to the CPS estimates of children within cells defined by race x age x sex.

b. $2^{nd}$ Stage Adjustment for Children

The overall second-stage adjustment procedure for children consists of the following steps.

STEP 1: Hispanic adjustment
STEP 2: Age adjustment

c. Second-Stage Adjustment for Adults

For adults, the following steps are employed in second-stage adjustment.

STEP 1: Raking procedure (all adults)
STEP 2: Hispanic adjustment
STEP 3: Raking procedure (all adults)
STEP 4: Hispanic adjustment
STEP 5: Raking procedure (non-Hispanic)

d. Final Longitudinal Weights

The final longitudinal weight (FW) for each person is equal to the weight generated after the second stage adjustment: FW = $W_i$ x $F_{1c}$ x $F_{2c}$, x (First-stage ratio) x $g_{ij}$

VII. DISCUSSION

Below we raise specific issues concerning SIPP longitudinal estimation.

A. Nonresponse is a particularly serious problem for a longitudinal survey such as SIPP since cumulative nonresponse increases as the life of the panel increases. A study on nonresponse behavior in SIPP has identified groups with differential nonresponse in the survey [12]. The effectiveness of the estimation procedure described above in reducing bias due to nonresponse is unknown. Research needs to be conducted to evaluate the effectiveness of this procedure in reducing these biases. If necessary, alternative adjustment methods should be explored.

B. Research in other areas of estimation need to be conducted. In particular:

1. Kalton, Lepkowski and Lin in [10] have examined the weighting adjustment versus imputation issue for handling nonresponse. They suggest that a combination of the two may be appropriate for certain types of wave nonresponse. Research needs to continue to determine wave nonresponse patterns that should be adjusted for by an imputation approach instead of a weighting adjustment.

2. The method described in this paper does not make use of those persons who failed to respond to the first interview but responded to subsequent interviews. The reliability of SIPP longitudinal estimates would improve if data on these persons could be utilized in an estimation procedure. Thus, we need to explore ways to use the data on such persons.

3. Longitudinal imputation in SIPP may adversely affect transition and spell estimates. Research needs to be conducted in this area to determine the effect of imputation on these type of estimates.

C. It is well known that a time-in-sample bias exists for other Census Bureau demographic surveys [8], [13]. Such a bias is likely to exist in SIPP. We have been unable to evaluate this bias in SIPP because of lack of data. As data accumulates for more SIPP panels, this bias should be evaluated.

D. There are other sources of biases in SIPP. For example, as respondents learn more about the survey their response to certain questions may be affected. Due to lack of knowledge on this and other biases, the procedure in this paper does not attempt to adjust for such biases. The effect of these biases need to be examined as they may affect SIPP estimates such as transition estimates. Research is now in progress to develop estimators with smaller bias for such estimates. A large scale effort in this area is needed. The research to be conducted should identify estimates with large biases as well as how to adjust for such biases.

## REFERENCES


[1] U.S. Department of Commerce, Bureau of the Census. "The Current Population Survey: Design and Methodology," by Robert Hanson. Technical Paper 40. Superintendent of Documents No.: C2.212:40; January 1978. Reprinted July, 1985.

[2] Census Bureau memorandum from G. Shapiro for T. O'Reagan, "Sampling Specifications for the 1984 Panel of the Survey of Income and Program Participation (SIPP)," January 31, 1983.

[3] Census Bureau memorandum from D. Hubble for Documentation, "SIPP 1984 Sample Design: Selection of 1984 Panel SIPP Sample PSUs," February 28, 1983.

[4] Nelson, D. McMillen, D., and Kasprzyk, D., "An Overview of the Survey of Income and Program Participation," SIPP Working Paper Series No. 8401, Update. U.S. Bureau of the Census, Washington, D.C., 1985.

[5] Census Bureau memorandum from C. Jones for T. Walsh, "SIPP 85: Cross-Sectional Weighting Specifications for Wave 1 -Revision," November 21, 1985.

[6] Census Bureau memorandum from C. Jones for T. Walsh, "SIPP 1984 - Specifications for Longitudinal Weighting of Persons," July 17, 1986.

[7] Ireland, C.T. and Kullback, S. "Contingency Tables with Given Marginals", Biometrika, vol. 55, 179-188, 1969.

[8] Census Bureau memorandum from R. Singh to G. Shapiro, "Nonsampling Errors in the National Crime Survey (NCS)," February 15, 1985.

[9] Judkins, D.R., Hubble, D. L., Dorsch, J. A., McMillen, D. B., Ernst, L. R., "Weighting of persons for SIPP Longitudinal Tabulation"; Proceedings of the Section on Survey Research Methods, American Statistical Association, 676-680, 1984.

[10] Kalton, G., Lepkowski, J., Lin, T., "Compensating for Wave Nonresponse in the 1979 ISDP Research Panel", Proceedings of the Section on Survey Research Methods, American Statistical Association, 372-377, 1985.

[11] Little, R. and David, M., "Weighting Adjustment for Nonresponse in Panel Surveys," Bureau of the Census Working Paper, 1983.

[12] Census Bureau memorandum from McArthur and Short for distribution list, "Measurement of Attrition from the SIPP through the Fifth Wave of 1984 Panel," April 10, 1986.

[13] U.S. Department of Commerce, Bureau of the Census, "An Error Profile: Employment as Measured by the Current Population Survey" by Camilla A. Brook and Barbara Bailar, Statistical Policy Working Paper 3., September 1973.


## ACKNOWLEDGEMENTS


David Judkins did much of the planning and early work upon which this paper is based. This paper owes much to Mr. Judkins' excellent insight. We also wish to thank Graham Kalton, Martin David and Daniel Horwitz for their contributions in helping us resolve issues related to SIPP longitudinal weighting.

The authors also wish to thank Joan George for her excellent typing and good humor through numerous revisions.