WEIGHTING OF PERSONS FOR SIPP LONGITUDINAL TABULATIONS

David Judkins, David Hubble, James Dorsch, David McMillen and Lawrence Ernst
U.S. Bureau Of the Census

I. INTRODUCTION

Since October of 1983, the Census Bureau has been conducting interviews for a new survey, the Survey of Income and Program Participation (SIPP). The survey will effect long-sought improvements in the measurement of annual income and the complex relationships between income flows, labor force participation, participation in government programs such as welfare, and tax policy. One of the products of the interviewing will be a set of longitudinal records on a probability sample of the population. The subject we address in this paper is the weighting of these longitudinal records so that the data may be analyzed.

We are aware of only two precedents for this weighting. They are the National Medical Care Expenditure Survey (NMCES) and the National Medical Care Utilization and Expenditure Survey (NMCUES). The latter was conducted jointly by the Research Triangle Institute and the National Opinion Research Center[2]. Some work was done on the problem for the Income Survey Development Program (ISDP)[6], but it was not implemented. The techniques used by them are among those under consideration for SIPP. Naturally though, we are also considering some new ideas. These ideas are still in a very preliminary form. We are presenting them here to get early reaction and suggestions from the statistical community.

Our general approach consists of three major steps. The first step is to derive an unbiased weight for each longitudinal record. This is not as straightforward as it seems due to the fact that a slightly different set of people is being interviewed each month. Section III discusses this step.

The second step is to make adjustments for those records that are incomplete. We will use imputation when part of an interview is missing. (See Samuhel's paper in this session [3].) We will also probably use imputation when a whole interview is missing where the missing interview is bracketed by good interviews. Our research on adjusting for records with more than one missing interview is in too preliminary a stage to report on. (One proposal has been made by Little and David[4].)

The third step is to correct for disproportional representation of demographic types to reduce variance and gain some consistency with the Current Population Survey (CPS). Section IV discusses this step.

Before discussing the weighting, it is essential that we define which of the many possible longitudinal universes is the universe for which estimates are to be provided. Section II deals with this problem.

Finally, we mention some of the important features of the design of SIPP. For more details, the reader is encouraged to first read an overview of the survey [5]. Roughly 20,000 households were interviewed between October 1983 and January 1984, inclusively. That set of interviews constitutes the first wave of the 1984 panel of SIPP. The Census Bureau will try to interview the persons in those households an additional seven or eight times in four-month waves, even if they move. We will also interview any persons who "usually reside" with anyone in the original cross-section for at least one-half of a calendar month. This extra interviewing will only be conducted for the time period that the joint residence is maintained. Only the original cross-section is followed through moves.

II. DEFINING THE LONGITUDINAL UNIVERSE OF PERSONS FOR SIPP

The SIPP universe at the beginning of any panel is persons who are members of the civilian non-institutional population, and members of the military not living in barracks on bases. Defining the longitudinal universe is somewhat more complicated. We begin by defining the possible ways persons can enter and exit this universe. Next we discuss the relationship between the cross-sectional universes and the longitudinal universe. The third topic of this section addresses the definition of table universes, and a discussion of calculating annual income for persons in the longitudinal universe.

There are two ways persons can enter the SIPP universe: 1) persons can move from overseas (immigrate or return), institutions, or from military barracks; 2) persons can be born to members of the universe.

Similarly, there are two methods of exiting the universe; 1) moving overseas, to an institution, or to military barracks 2) dying. Given these conditions of entering and exiting the universe, and a definition of the initial universe, we can define the universe at any subsequent point in time, and the means by which the universe grows and diminishes over time. The next problem is to make the transition from the cross-sectional universes to a single longitudinal universe.

There are three methods of defining a longitudinal universe: 1) the composition can be fixed at some point in time; 2) the universe may be defined as the union of some set of cross-sectional universes; and 3) the universe may be defined as the intersection of some set of cross-sectional universes.

A longitudinal universe may be defined at a given point in time. For example, we can take the civilian noninstitutional population at the time the sample is drawn, at the midpoint of the panel duration, or at the end of the panel to define the universe of interest. Of course, the time point chosen could be any time point within the duration of the panel. This rather narrow definition of the universe has an advantage in its simplicity, but also several disadvantages. Dependent on the chosen point in time, this definition produces a strictly declining population, a first increasing and then decreasing population, or a strictly

increasing population. In the first case all entrants are excluded from the longitudinal universe, and only exits are allowed to alter the universe. In the second case, entry is allowed and exit is denied until the midpoint, when the situation reverses. In the last case, all those who exit during the panel are excluded from the longitudinal universe and only entries are allowed to alter the universe. In addition, it is difficult to argue why one point or another should be chosen as the point in time to define the universe, and for some purposes you may need a different point than the one originally chosen.

The next two definitions build from the above idea that a universe may be defined at any point during the panel. Let us assume then a set of universes each defined at a different point in time. To further simplify discussion, let us assume a set of twelve monthly universes defined at the midpoint of each month. The two options are to use either the union or the intersection of these sets.

Consider first the union of sets. The union of these monthly universes is all persons who were at some point during the year members of the civilian noninstitutional population. In other words, all members of the target population plus all persons who enter or exit during the year are included in the union of sets definition. This is the most inclusive of the universe definitions offered here, and the one which best captures the dynamic characteristics of the population. Some of the disadvantages of this type of definition will be raised below in the discussion of tabulations and table universes.

An alternative to the union of sets is the intersection of the set of twelve monthly cross-sectional universes. Here we include in the longitudinal universe only those persons who were members of all of the cross-sectional universes. In other words, only those persons who were members of the civilian noninstitutional population or the special military categories on the fifteenth of each of the twelve months. This definition is even more restricted than the point-in-time definition. This intersection of sets definition produces a static population. That is to say there is no entering or exiting allowed.

Of the three longitudinal definitions offered here, only the union of sets incorporates the dynamic qualities that are inherent in a longitudinal process.

That would seem to make it the logical choice; however, this is also the definition that produces the most complications when tabulating data. Consider, for example, tabulating marital status at the beginning of the year with marital status at the end of the year. There is no place in such a table for persons who were in universe at one point in time, and not in the universe at the other point in time. For the union of sets definition there is a need for both a column and a row for persons not in the universe at time 1 or not in universe at time 2. For those definitions that allow exiting only a column for persons not in the universe at time 2 is necessary as long as the beginning point of the universe and the tables are the same.

Similar problems arise in computing annual income. Aggregating across months is simple, but it is not clear how to compare income amounts for full year and part year persons. That is simply to say that a $6,000 income for 6 months and a $6,000 income for 12 months are not the same.

III. INITIAL WEIGHTING

For SIPP, as for ISDP, a cross-section of the population will be followed for a period of time. Data will also be collected on the people that the original cross-section live with. The original idea was that only the data on the people in the original cross-section would be used in person longitudinal tabulations; the data on the other people would be used only to provide the "household experience" of the original cross-section. We are now reexamining that idea. The data on the other people can be used to better understand the experience of new entrants to the SIPP universe. Furthermore, there are ways to use these data more intensively to gain valuable variance reductions. Unfortunately, these procedures require strong assumptions for unbiasedness. In the following sections, we explore the trade-off. We first discuss whether the data on the other people should be used. We then discuss how to construct weighting procedures that use these data more or less intensively.

A. Variance Reduction Versus Bias Control.

Let us first define some terms and clarify the type of parameters to be estimated. We divide the sample people into three groups. A person is an original sample person if he/she is a member of the original cross-section.[1] A person is an associated sample person if he/she was a member of the eligible population at the time the cross-section was selected but happened not to be selected. Anyone else is an additional sample person. This last group consists of recent discharges from institutions, new immigrants, and people moving out of military barracks. The type of parameter to be estimated is the frequency of some pattern of labor force participation, program participation, income receipt, etcetera, by demographic characteristics, housing characteristic, geographical unit, educational background, etcetera. A simple example is the frequency of women who were receiving public assistance in January 1984 but were not

1 A person in original cross-section of households who was 15 years old or older at the time of the first interview is definitely an original sample person. Twelve, thirteen, and fourteen year old children are more difficult to classify. At first, no questionnaires are filled out for them and they are not followed in the rare event of an unaccompanied move. However, after they turn 15, they are treated the same as any other original sample person. We will treat them here as original sample people. Children eleven or younger are not classified at all.

receiving it in December 1984.

The original idea was to estimate parameters like this one by summing the weights of all original sample persons with the desired characteristics. Data on associated and additional sample people are needed only to classify original sample people with respect to household characteristics; for example, was the original sample person living in a household in which at least one member received social security? Given this scheme, no data are needed on associated or additional sample people for the period that they don't reside with original sample people. Hence, we do not follow associated or additional sample people if they separate from original sample people. Clearly then, the data on associated and additional sample people are frequently incomplete.

Despite this incompleteness, we are now considering ways to squeeze more information out of this data. The first way is to provide estimates for the "union" universe using the data on additional sample. The second way is to use the data on both types to reduce variances. To begin the argument for this second use, we first point out that for shorter time periods these data are frequently either complete or nonexistent. (Throughout this section, by complete we mean complete ignoring nonresponse.) This is always true for 1 month periods, usually true for 3 month periods, and frequently true for 12 month periods. For example, suppose that Ruth is an original sample person interviewed in October 1983. In November, she marries Jack, who was in the October SIPP universe. They stay together at least through April 1985. Then Jack is an associated sample person on whom we have complete 1984 data. Alternatively, suppose that Jack was living in a military barracks in October 1983. Then he is an additional sample person on whom we have complete 1984 data. There will obviously be many more cases in these complete categories for 1985 data. Furthermore, there will be many cases where we are only missing one or two months of data.

Intuitively, it seems wasteful to give zero weights to these cases with complete or almost complete data, as originally intended. On the other hand, zero weights must be assigned to the seriously incomplete cases to avoid large-scale imputation. One possible solution is obtained by initially assigning strictly positive weights to all cases, including these that are incomplete due to field procedures, and then treating the incomplete cases as if they were caused by non-response. Imputation would be used for the almost complete cases. Note then that the seriously incomplete cases would have zero weights, while the other cases would have positive weights. If enough data has been collected on the associated and additional sample people to correctly model the probability of this type of nonresponse, then we would still have unbiased estimators.

An example of the type of model required is that starting from a given social-economic stratum, the new economic situation of a male divorcee does not depend on whether he or his ex-spouse was the original sample person. Here we stress that if a person has responded to even a single wave of SIPP, then we have an extraordinary wealth of data available for modeling.

Future Study

Of course, we will never know for certain whether such a model is correct. There is a risk of biasing the estimators, and as a rule the Bureau is willing to risk biases for decreases in variance only if there is some evidence that the bias squared is substantially less than the variance decrease. Our plans at this time are not well formulated. A reasonable first step is to quantify for each proposed weighting procedure the frequency of positively weighted incomplete cases by the severity of the incompleteness. The only source for this information is the ISDP. We are currently working on ways to get appropriate tabulations for it.

8. Construction of Unbiased Weighting Procedures

Below we present a very simple result that characterizes a general class of unbiased procedures. Reflection on this result quickly helps one to understand that there are infinitely many unbiased procedures. Most of them are totally inappropriate, but it is very possible that better and radically different weighting procedures exist than have yet been conceived.

Let $x = \sum_{i}^{N} x_i$ be the parameter of interest to be estimated where x_i is the value of the characteristic for the i^{th} unit. Let w_i be a random variable associated with the i^{th} unit such that $E(w_i) = 1$.

$$E(Y) = E(\sum_{i}^{N} w_i x_i) = \sum_{i}^{N} E(w_i) x_i = \sum_{i}^{N} x_i = x.$$

If the probability of selection is known for all units, it is common to take

$$w_i = \begin{cases} \text{inverse probability of selection if } i^{th} \text{ unit is in sample;} \\ 0 \text{ otherwise.} \end{cases}$$

This definition of w_i is not, however, necessary. In this case it is impossible since the probabilities are unknown.

Each sample person has a cross-sectional weight for every month that they are in the universe. These cross-sectional weights have expected value of unity, are strictly positive for the months that the person is in sample, and are zero for the months that the person is not in sample. By choosing the longitudinal weight to be the cross-sectional weight at a particular time or the average of the cross-sectional weights

at several points in time, we can construct longitudinal weighting procedures that use different subsets for the overall data set.

In this section we present four longitudinal weighting procedures for computing unbiased estimates for persons. They are all presented in terms of the "union" universe, but they can be easily modified for the "intersection" universe by assigning a zero weight to any person who is not in every one of the 12 cross-sectional universes. In Section III.C we compare the procedures with respect to the use of data collected on associated sample persons and additional sample persons. In the full paper there is an additional section with examples of the application of these procedures.

Procedure 1. Entry Date Weight (ED)

Each person receives a single longitudinal weight for any time interval that contains at least part of the period for which the person was in the universe, namely the cross-sectional weight for the person at his/her entry date into the universe. For all original and associated sample persons, the entry date into the universe is the start of the panel, so their longitudinal weights are their Wave 1 cross-sectional weights. For those who enter the universe after Wave 1, (additional sample persons), the longitudinal weight is the cross-sectional weight of the household, of which they are a member, as of the date they enter the universe. If the cross-sectional weight of the household at that date is zero, then the additional sample person's longitudinal weight is zero.

Procedure 2. Beginning Date of Time Interval Weight (BDI)

Each person receives a longitudinal weight valid for all time intervals with the same beginning date. Persons in the universe at the beginning date of the time interval are assigned their respective cross-sectional weights for that date. Persons that enter the universe during the time interval are assigned their respective cross-sectional weights as of the date they enter it, as in Procedure 1.

Procedure 3. "Mid" Date of the Time Interval Weight (MDI)

This procedure is similar to Procedure 2. Each person receives a longitudinal weight valid for a specific time interval. Persons in the universe at the "mid" date of the time interval are assigned their respective cross-sectional weights at that date. The difference is that instead of the person longitudinal weights being determined at the beginning date of the time interval, these weights are determined at some predesignated date within the time interval. Persons that enter the universe during the time interval but after the mid date are assigned their respective cross-sectional weights as of the date they enter it, as in Procedure 1 and 2. Persons who leave the universe before the "mid" date are assigned their respective cross-sectional weights as of the date they leave it.

Procedure 4. Average Cross-Sectional Weight (ACS)

Each person recieves a longitudinal weight valid for a specific time interval. Persons that remain in the universe throughout the interval are assigned the average of their respective monthly cross-sectional weights. Persons that enter or leave the universe are assigned the average of their respective monthly cross-sectional weights for the months they were in the universe during the time interval. Positive weights are assigned to all sample persons. A more formal definition is given below.

Let U_i = number of months the i^{th} person was in the universe during the specified time interval

Let C_i = sum of the monthly cross-sectional weights of the i^{th} person in the specified time interval

Then the person longitudinal weight is C_i/U_i.

C. Comparison of Procedures

In this section we describe in detail the types of complete and incomplete cases that are used by each procedure. First, we need to define some notation. Let

t_B = the first month that a person is in the universe,

t_E = the last month that a person is in the universe,

t_1 = the first month that a person is in sample,

t_2 = the last month that a person is in sample,

t_m = the mid-month of the interval of interest.

The description is given in Table 1. The first 14 cases comprise the "intersection" universe. The remaining 32 cases fill out the "union" universe. Each case is marked as having complete, partial or no data for the interval of interest. Of course, all of this is assuming perfect response. The only type of missingness that we are discussing here is that caused by operational procedures. On the right, there is a column for each procedure with an "X" if the procedure uses the case.

The entry date procedure uses the perfect cases 1,15,17, and 18, but does not use the perfect cases 2 and 16; the partial cases 3,5, and 19-27; and cases 12 and 44 for which no relevant data exists. The beginning date of interval and mid date of interval procedures both use all of the perfect cases, more of the partial cases and none of the completely missing cases. We thus think that these two procedures will tend to yield smaller variances than the entry date procedure with possibly some small increase in the risk of bias. The average cross-sectional procedure is the most aggressive in utilizing partial data. It uses all the perfect and partial cases and none of the completely missing cases.

Also note that it assigns smaller weights, in general, to the partial cases than the perfect cases. We think it will tend to yield the smallest variances with the greatest risk of bias.

IV. CONTROLS

We are currently considering the adjustment of SIPP longitudinal weights so as to achieve the variance reductions associated with ratio estimation while also causing agreement with SIPP cross-sectional controls on a monthly basis; i.e., in addition to simple undercoverage adjustments we are considering the possibility of forcing the sum of the longitudinal weights of all persons in the universe in a given month to equal the cross-sectional population control for that month. Since longitudinal weights are fixed over time while the universe fluctuates over time, such agreement will not occur unless proper steps are taken to ensure it. We are also considering adjustments to force spouses to have equal longitudinal weights. We are considering these two possibilities in order to enhance the face validity of the survey at the least possible cost of reduced precision.

Objectives

The primary reason for ratio adjustment of longitudinal weights is to reduce variances of longitudinal weights by ensuring representativeness with respect to demographic variables which are highly correlated with the variables to be measured. (This is frequently referred to as post-stratification.) To the extent that it corrects for differential undercoverage, it is also hoped that bias is reduced by ratio adjustment.

A reasonably good adjustment is to proportionately adjust the weights of persons by demographic type in a specified month so that the weighted counts agree with independent population estimates by demographic type for that month. Persons not in sample in the chosen month are assigned the factor for their demographic type. This approach operates under the assumption that the degree to which the sample represents each demographic type is not highly variable over time. This adjustment does not adjust weights to monthly controls other than those for the chosen month. Another approach is to make the adjustment for all persons for each of the 12 data months, then assign to a person the average of the 12 factors for his/her cell. Such an adjustment would tend to be influenced less by the vagaries of sample selection.

Addressed here is the more complex problem of adjusting weights for disproportional representation in a manner such that consistency with cross-sectional controls is achieved for each month. This problem has a multitude of solutions. However, the solution we seek should be the one which provides the greatest variance reduction. One possible solution is to first adjust weights as outlined in the above paragraph, then further adjust them so that the desired monthly consistency is achieved while minimizing the amount by which weights are further adjusted. This can be done with Lagrange multipliers or with linear programming. This approach preserves the benefits of the initial adjustment by demographic variables provided that this second adjustment causes relatively small changes in weights. Research is needed to determine whether the second adjustment would indeed cause only small changes.

A further refinement would be to adjust so that spouses have equal weights. Naturally, persons undergo changes in marital status during the year; some persons may have more than one spouse over a one year period. Define a "marriage group" to be a group of persons in the SIPP sample, each of whom has been or is married to at least one other person in the group during the data year. It is possible to perform an adjustment so that all persons in a given marriage group have equal weights. This last adjustment would cause slight disagreements between longitudinal population estimates and monthly controls; it appears likely that such disagreements could be made arbitrarily small by iteratively repeating the two adjustment steps for consistency with cross-sectional estimates and consistency within marriage groups. For more details, see our full paper.

REFERENCES

[1] Census Bureau memorandum from C. Jones to T. Walsh, "Cross-Sectional Weighting Specifications for the First Wave of the 1984 Panel of the Survey of Income and Program Participation (SIPP)," November 25, 1983.

[2] Jones, Bruce L., "Development of Sample Weights for the National Household Survey Component of the National Medical Care Utilization and Medicare Utilization and Expenditure Survey," April 1982.

[3] Samuhel, Michael E., "Longitudinal Item Imputation in a Complex Survey," presented to the Survey Research Methods Section of the American Statistical Association during the 1984 Annual Meetings.

[4] Little, R.J.A. and David, M., "Weighting Adjustments for Nonresponse in Panel Surveys," 1983 Working Paper.

[5] Nelson, D., McMillen, D., and Kasprzyk, D., "An Overview of the Survey of Income and Program Participation," SIPP Working Paper Series No. 8401. U.S. Bureau of the Census, Washington, D.C. 1984.

[6] Kasprzyk, D. and Kalton, G., "Longitudinal Weighting in the Income Survey Development Program," in Technical, Conceptual and Administrative Lessons of the Income Survey Development Program (ISDP), Papers presented at a conference, October 6-7, 1982. Social Science Research Council, Washington D.C., 1983.

[7] Sirken, Monroe G., "Household Surveys with Multiplicity," Journal of the American Statistical Association, 65, No. 329 (1970), 257-66.

Table 1. Case Utilization by Procedure

Case	Preceeding Time Interval	Interval of Interest	Succeeding Time Interval	Completeness	Procedure ED	BDI	MDI	ACS
1	$t_B=t_1$		$t_2 \leq t_E$	Perfect	X	X	X	X
2	$t_B<t_1$		$t_2 \leq t_E$	"		X	X	X
3	$t_B=t_1$	$t_m \leq t_2$	t_E	Partial	X	X	X	X
4	$t_B<t_1$	$t_m \leq t_2$	t_E	"		X	X	X
5	$t_B=t_1$	$t_2<t_m$	t_E	"	X	X		X
6	$t_B<t_1$	$t_2<t_m$	t_E	"		X		X
7	t_B	$t_1 \leq t_m$	$t_2 \leq t_E$	"			X	X
8	t_B	$t_1 \leq t_m \leq t_2$	t_E	"			X	X
9	t_B	$t_1 \leq t_2 < t_m$	t_E	"				X
10	t_B	$t_m < t_1 \leq t_2$	t_E	"				X
11	t_B	$t_m < t_1$	$t_2 \leq t_E$	"				X
12	$t_B=t_1 \leq t_2$		t_E	No Data	X			
13	$t_B<t_1 \leq t_2$		t_E	"				
14	t_B		$t_1 \leq t_2 \leq t_E$ z	"				
15	$t_B=t_1$	$t_2=t_E$		Perfect	X	X	X	X
16	$t_B<t_1$	$t_2=t_E$		"		X	X	X
17		$t_B=t_1$	$t_2 \leq t_E$	"	X	X	X	X
18		$t_B=t_1$ and $t_2=t_E$		"	X	X	X	X
19	$t_B=t_1$	$t_m \leq t_2 < t_E$		Partial	X	X	X	X
20		$t_B=t_1$	$t_2<t_E$	"	X	X	X	X
21		$t_B=t_1$ and $t_m \leq t_2$	t_E	"	X	X	X	X
22		$t_B=t_1$ and $t_m \leq t_2 < t_E$		"	X	X	X	X
23	$t_B=t_1$	$t_2<t_m \leq t_E$		"	X	X		X
24	$t_B=t_1$	$t_2<t_E<t_m$		"	X	X		X
25		$t_B=t_1 \leq t_2<t_m$	t_E	"	X	X		X
26		$t_B=t_1 \leq t_2<t_m \leq t_E$		"	X	X		X
27		$t_B=t_1 \leq t_2<t_E \leq t_m$		"	X	X		X
28	$t_B<t_1$	$t_m \leq t_2 < t_E$		"		X	X	X
29	$t_B<t_1$	$t_2<t_m \leq t_E$		"		X		X
30	$t_B<t_1$	$t_2<t_E \leq t_m$		"		X		X
31	t_B	$t_1 \leq t_m \leq t_2 < t_E$		Partial			X	X
32	t_B	$t_1 \leq t_m$ and $t_2=t_E$		"			X	X
33		$t_B<t_1 \leq t_m$	$t_2 \leq t_E$	"			X	X
34		$t_B<t_1 \leq t_m \leq t_2$	t_E	"			X	X
35		$t_B<t_1 \leq t_m \leq t_2 \leq t_E$		"			X	X
36	t_B	$t_1 \leq t_2<t_m \leq t_E$		"				X
37	t_B	$t_1 \leq t_2<t_E \leq t_m$		"				X
38	t_B	$t_m<t_1 \leq t_2 \leq t_E$		"				X
39		$t_B<t_1 \leq t_2<t_m$	t_E	"				X
40		$t_m<t_B<t_1$	$t_2 \leq t_E$	"				X
41		$t_m<t_B<t_1 \leq t_2$	t_E	"				X
42		$t_B<t_1 \leq t_2<t_E \leq t_m$		"				X
43		$t_m \leq t_B<t_1 \leq t_2 \leq t_E$		"				X
44	$t_B=t_1 \leq t_2$	t_E		No Data	X			
45	$t_B<t_1 \leq t_2$	t_E		"				
46		t_B	$t_1 \leq t_2 \leq t_E$	"				